SEMANTIC MANIPULATION AND BUSINESS CONTEXT IN BIG DATA ANALYTICS

Loan Thi Ngoc Dinh

A dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Under the Supervision of:

Dr Gour Karmakar, A/Prof Joarder Kamruzzaman, A/Prof Andrew Stranieri

School of Science, Engineering and Information Technology

Federation University Australia, Gippsland Campus

August 2018

Declaration

I declare that this dissertation is my original work and I have not knowingly added copyright content to my work without the owner's permission.

© Loan Thi Ngoc Dinh (2019).

Acknowledgements

I would like to express my profound gratitude and greatest appreciation to my main supervisor, Dr Gour Karmakar, for his tireless efforts in ensuring the quality of my research, his encouragement, endless patience and energetic supervision. Without his help and support, this research would not have been completed.

I especially thank my associate supervisor, A/Prof Joarder Kamruzzaman, for his professional guidance and encouragement.

I also thank A/Prof Andrew Stranieri for his valuable suggestions and support, and Dr Alex McKnight for proofreading the final draft for grammatical and stylistic errors.

I would like to thank Federation University Australia and the Vietnamese Government for giving me the opportunity to carry out my research in an excellent environment. I also thank them for providing me with the necessary resources, materials, financial support and scholarships.

I am also thankful to all staff and my best friend Dr Priyabrata Karmakar at Gippsland Campus for their cordial support and inspiration. Discussions with them have enriched my conceptions about and knowledge of this research work.

Finally, my heartiest love and gratitude are due to my family, especially my parents, for their endless love, inspiration, encourage and support.

Abstract

Business organisations receive a huge amount of data from many sources every day. These data are known as big data. Since they are mostly unstructured, big data creates a complex problem of how to capture, manage, analyse and then derive meaningful information from them.

To deal with the challenges that big data has brought, this research proposes a new technique in big data analytics in the business area to integrate semantically meaningful information relevant to textual queries and business context. To achieve this aim, this study makes three major related contributions. Firstly, the relationship between business processes and strategies is established using the concept of a rule-based inference model via facts and annotations. This relationship is required to determine the importance of a big data query for a business organisation. Secondly, we introduce approaches to determine the significance level of a query, by incorporating the process-strategy relationship, process contributions and priority of business strategies. Thirdly, the proposed data analytic technique embeds business context into the bedrock of data collection and analysis process.

The first two contributions were implemented using Python programming language including the Pyke package (Pyke is built in the Python environment and has an artificial intelligence tool for the development of expert systems) and their performances were analysed based on a business use case. The last contribution was implemented mainly in the Hadoop and Java programs.

Results show that the first contribution successfully establishes the process-strategy relationship, the second calculates the significance level of a query in relation to a business organisation, while the third reveals the huge impact of query significance level and business context on big data collection and captures deep business insights.

Abbreviations

CRISP-DM	Cross-Industry Standard Process for Data Mining
BI	Business Intelligence
BIAA	Business Intelligence And Analytics
B2B	Business-To-Business
BPM	Business Process Management
BP	Business Process
BS	Business Strategy
IR	Information Retrieval
DSS	Decision Support System
OLAP	Online Analytical Processing
ETL	Extract-Transform-Load
HDFS	Hadoop Distributed File System
KFB	Knowledge Face Base
SLQBP	Significance Level of Query considering Business Processes
SLQBPS	Significance Level of Query considering Business Processes and Business Strategies

Notations

$\boldsymbol{k}$	Keyword
$\boldsymbol{w}$	Semantic word
$\boldsymbol{\xi(k,w)}$	Semantic similarity between keyword k and semantic word w
$\boldsymbol{p}$	Business process
$\boldsymbol{q}$	Query
$\boldsymbol{\xi_{qp}}$	Semantic similarity between business process p and query q
$\boldsymbol{\xi_q}$	Semantic similarity between the query and all business processes
$\boldsymbol{s}$	Business strategy
$\boldsymbol{\bar{\xi}_q}$	Median value of ξ_q for all core p
$\boldsymbol{\tilde{\xi}_q}$	Median value of ξ_q for all non-core p
$\boldsymbol{n}$	Number of keywords in query
$\boldsymbol{a}$	Annotation of business process
$\boldsymbol{W_p^s}$	Contribution of business process p to business strategy s
$\boldsymbol{W_s}$	Priority of business strategy
$\boldsymbol{W_p^f}$	Final contribution of business process
$\boldsymbol{W_n}$	Weight for non-core p
$\boldsymbol{z}$	Total number of strategies considered in the case study

y_p	Number of strategies in which process p contributes to
c	Sensitivity of process contribution with respect to number of associated strategies
ϵ_q	Significance level of a query based on SLQBP
ϵ_q^s	Significance level of a query based on SLQBPS

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Big data

Big data encompasses a wide variety of data types, including text, image, audio and video, which are continuously generated from diverse sources such as the web, social media, mobile applications, sensor devices, networks, and data storage. Big data is also generated by an organisation itself. Because of the explosive growth in the volume of global data in recent times, many enterprises, not only large but also medium- and small-sized, are aware of the importance of extracting key information from big data. This can be seen not only in the way businesses are investing money in buying new technologies to extract precious information from big data, but also in their efforts to accelerate big data research and applications [3–5].

Big data is divided into four main categories: (i) structured (ii) unstructured (iii) semi-structured and (iv) mixed [5], as presented in Table 1.1. According to Marcos et al. [5], a major proportion of data produced today is either unstructured or semi-structured. Traditional database management systems deal with structured data and therefore cannot manipulate unstructured and other data types efficiently (e.g., email, documents, social network contents), or sensor data (such as images, audio, video and other data types). Therefore, in parallel with database management systems, a number of big data analysis methods are being developed to deal with unstructured or semi-structured data.

There are seven main high-level dimensions of big data: (i) variety, (ii) velocity, (iii) volume, (iv) veracity, (v) validity, (vi) volatility and (vii) value, as shown in Table 1.2 [6]. The two main goals for the development of effective methods to analyse such high-dimensional data are to: (i) accurately predict future observations to make informed decisions and (ii) gain insight into the relationships among the various features of the data and their impacts for many purposes in many application domains from health and science to business.

Table 1.1: Types of big data

Data type	Description
Structured	Formal schema and data models
Unstructured	No pre-defined data model
Semi-structured	Lacks strict data model structure
Mixed	Various types together

Table 1.2: Seven dimensions of big data

Dimension	Description
Variety	Data types
Velocity	Data production and processing speed
Volume	Data size
Veracity	Data reliability and trust

Validity	Data correctness and accuracy with respect to the intended usage
Volatility	Big data is volatile and destroyed when retention policies or warranty periods expire.
Value	Worth derived from exploiting big data

However, the inherent complexity of big data derived from many sources poses a major challenge for analytics. For example, high dimensionality brings noise accumulation, spurious correlations and incidental homogeneity. In addition, high dimensionality combined with large sample size creates issues such as high computational cost and algorithmic instability. The massive samples in big data are typically aggregated from multiple sources at different time points using different technologies [5,7], further contributing to the complexity.

According to a number of studies, considering contextual information during data processing can deal with the issues mentioned above [8–11]. For example, the study conducted by Aknouche et al. [12] on the perception gap in collecting relevant information, the authors emphasised that most of the information retrieval systems rely on the retrieval decisions made by queries and document collection processes. However, since the search context is ignored, a large number of irrelevant results occur in these systems [12].

Since semantic information manipulation can effectively capture deep insights, it is one of the main research aims of this thesis. Therefore, before investigating existing context-based big data analytic approaches, an overview of semantic big data analytic techniques is presented in the next section.

1.1.2 Semantic-based big data analytic techniques

In the present information era, most organisations not only want to collect data but also extract the meaning, importance and value of the data so that they can use it in their decision-making processes. According to [13], the term "data analytics" refers to a process of applying algorithms with the purpose of analysing sets of data and extracting useful information and unknown patterns, and their relationships. In data analytics, in addition to obtaining useful information, valid but hidden patterns are also extracted from enormous data sets and the key relations among the stored variables are discovered. In another study, Song and Kusiak [14] pointed out that technologies and research have a big influence on big data analytics. Numerous decision makers have become extremely interested in learning from earlier data, and consequently, significant competitive advantages have been gained.

Most traditional data analytic techniques are not suitable for the analysis of such a huge variety of data [15]. Therefore, there is a requirement for more effective and efficient ways of processing big data. Sinoara et al. [16] emphasised the impact of semantic information on document analysis, as the utilisation of similar vocabulary may introduce diverse thoughts regarding similar subjects.

Embedding semantic information into big data anlytics has attracted the attention of a number of research communities, especially in the use of semantic information for text mining and sentiment analysis .

Text mining Text mining is used to analyse a document or a set of documents with the intention of understanding the contents and the meaning of the information they contain. Today, a great deal of information is stored as text. Therefore, text mining has become very important. In [17], Sanchez et al. emphasized that, while data mining deals with structured data, text presents special characteristics that basically follow a non-relational pattern. However, the methods used in text mining are not suitable for analysing other types of information such as images, audio and video.

Sentiment analysis: Sentiment analysis is also known as opinion mining [18]. The aim of sentiment analysis is to focus on analysing, capturing and understanding the feelings, emotions and reactions based on the writers attitude to a particular topic or product as positive, negative or neutral, and this is assisted by text mining. There are two main components are included in sentiment analysis: (1) text analytics and (2) natural language processing [19]. Sentiment analysis identifies and extracts information by finding the key words that are relevant to a sentiment, in addition to relations among words, so that sentiments can be captured accurately [18]. Since people have become much more active on online social networks these days, they spend a considerable amount of time online, and prefer working and entertaining themselves via the Internet, such as shopping online, chatting and watching movies. Therefore, online social media are currently becoming much more important, for example, for opinion dynamics via blogs, forums, product reviews and social data from many sources of social media sites such as Facebook and Twitter.

1.1.3 Context-based big data analytics

Most big data analytic methods have been developed based on traditional methods and do not consider contextual information [20]. As a result, these lead to information related to actual users but since the search context is ignored, these approaches produce great numbers of irrelevant results [21].

In general, all current data mining methods face great difficulties because they are expected to be able to handle many issues of big data, including its unprecedented heterogeneity, speed, accuracy, volume, privacy and trust, but they were not initially designed with these issues in mind. According to [22,23], a large number of research projects have been undertaken in order to improve the existing methods and techniques and overcome the challenges in big data in many ways, such as by applying massive parallel processing architectures and novel distributed storage systems, and designing innovative mining techniques based on new frameworks/platforms (e.g., Hadoop [22], Mahout [24], Giraph [25]) with the potential to successfully overcome the aforementioned challenges and reshape the future of data mining technology.

On the other hand, a number of big data analytics methods which are applied to business intelligence and analytics (BIAA) are reviewed in [26]. The evolution of these methods can be divided into three generations. In the first generation of BIAA, most methods were developed to analyse database management systems based on structured content. In the second generation, unlike the first generation, this is an integration of mature and scalable techniques in analysing web-based and unstructured content. In the third generation, the big data analytic methods currently focus on analysing mobile and sensor-based content. As a result, it can be seen that data analytic techniques have been developed based on the growing emphasis on unstructured data. The focus on query in BIAA has recently shifted more to context awareness [26], unlike traditional methods that are non-context-based and focus on individual queries lacking exact context. To overcome this issue, a context-based query classification based on neighbouring queries and their corresponding clicked web pages has been considered [27, 28]. To achieve this, context information is incorporated into the problem of query classification by using conditional random field models [29, 30]. User context plays a vital role in information retrieval and to date no information retrieval system exists which takes into account all intentions and perceptions of users. In [31], information retrieval techniques were proposed by integrating both user context and query context based on language modelling. The query context includes the integration of linguistic and semantic knowledge about the user query (e.g., instead of explicitly assigning meaning to a word, distinctions between word usages are made based on different contexts, and context words are added to relations in order to exploit another relevant word context within the query) and user context can be found either by the users domain of interest or the topic of interest. For example, if a user mentions "Apple" in a query, there is difficulty in recognizing if he/she is interested in fruit or a technology company. Many existing methodologies classify the query into both categories "Fruit" and "Technology Company" without understanding the users search intent. However, by considering the users domain of interest or the topic of interest, if a query "Fresh" can be found before "Apple", the category of "Fruit" will be assigned to the user interest. In contrast, if the user issues some queries related to technological items such as "laptop" and "computer" before "Apple", the users interest will be "Technology Company".

The capture of data and their values relevant to a business system demands the integration of its context in big data analytics. This is because the utilisation of context in other application domains, such as information retrieval [12,32–34], business process and workflow [9–11], and reasoning scenarios [35], has delivered significant benefits.

As indicated previously, since the consideration of context (e.g., user context, query context) and its relation to big data analysis has been the consistent focus of the data science community, a number of existing data analytic methods consider contextual information that includes, but is not limited to, information retrieval considering the context of a query and its users [8], query-driven context-aware recommendations [36] and so on. However, these systems still have many challenges [3,37] and none of them have considered the context of a business in either data collection or analysis. In contrast, it is clearly evident from the research in big data analytics that, if the business context is embedded into both data collection and analysis, it is possible to capture more and deeper insights into the data directly relevant to a business more accurately. This will help a business to develop a more feasible and achievable business strategy.

1.2 Motivation

Capturing deep insights from data is not only a major challenge for business organisations but is also a major issue for the relevant research community. Since big data has a wide variety of types and comes from diverse sources, a huge amount of data is generated by both external (e.g., internet) and internal sources. To capture valuable information to reaching any business goal is a demanding and challenging research problem. Moreover, it is also difficult to collect the information that is really valuable and relevant to business needs from this huge amount of big data. For this reason, extracting the real key value information for meeting any business analytics need is also generally poor.

As a consequence of the variety of data from diverse sources, most data types are semi-structured or unstructured. These data are in large amounts and it takes an excessively long time if a business wants to collect all data. Moreover, data analytic methods have very high computational complexity to extract meaningful and semantic information and hidden clues from the huge variety and volume of semi-structured and unstructured data. This makes them either unsuitable or less suitable for real-time or near real-time applications. To address this issue, it is crucial to scale the semantic information dynamically for use in big data collection and analytic processes, considering their significance from a business operational and strategic decision-making perspective.

There are many traditional methods of data analytics but most are currently only suitable for analysing structured data and not as effective for semi/unstructured data. In addition, these methods still have many challenges and none consider the context of a business in order to capture deep insights into data directly relevant to a business organisation in either the data collection or analysis processes. Therefore, business organisations need to adopt filtering mechanisms so that they can collect those data that are really relevant to the business perspective and analyse them according to their requirements. This demands the consideration of business context in big data gathering and analysis. Furthermore, this imposes a significant challenge to the data science community to manage and provide smart and instant data analytics for a business for informed decision-making.

1.3 Research Objectives

Based on the research issues identified in the previous section, this dissertation aims to address the following research objectives:

Objective 1: To investigate the development of a technique that can be used in dynamically scaling the level of semantic information used in both textual big data collection and analytic processes, depending on their importance from a business perspective.
Objective 2: To capture deep insights into data which are directly relevant to a business organisation, we aim to explore the business contextual information that can be used in both data collection and analytics.
Objective 3: To introduce textual big data query-based data collection and

analytical techniques which can incorporate and dynamically scale semantic information, as mentioned in Objective 1 and embed the business contextual information in the bedrock of both data collection and analytics (Objective 2).

1.4 Overview Of Contributions

An overview of the major contributions of this thesis is given below:

(i) Development of a rule-based inference model depending on facts of business strategies and annotations of business processes to establish dynamically the relationship between business strategies and processes. This presents a significant component of a research project required to address research Objective 1 and has been published in [38]

(ii) Calculation of the significance level of textual queries to scale semantic information during both data collection and analytics to present a vehicle to produce a solution to research Objective 1. For this research project, we have developed two approaches. The first approach calculates the significance level of queries based on the intuitively selected weight of business processes published in [39], while the second determines the significance level of a query by exploiting the contributions of business processes and the priorities of business strategies [40].

(iii) Embedding contextual information into big data analytics to capture data directly relevant to a business organisation. This directly addresses research Objective 2. The outcome of the preliminary research related to this topic has been published in [41]. Note that the business context may be represented in many different ways depending on the business application requirements. In this thesis, emphasis is given to customer feedback analysis and separating the business context from semantic information manipulation. Therefore, we represent the business context of an organisation in an innovative way, using its main slogans utilised to attract more customers for generating more revenue.

(iv) Implementation of a technique that allows scaling semantic information and also embedding business contextual information during both data collection and analytics (Objective 3). The technique is implemented in the Hadoop framework using a textual big data query relevant to a grocery shop. The performance of this technique has been evaluated considering a set of semantic keywords relevant to query keywords selected by the use of the significance level of that query and the different sets of keywords derived from the slogans representing the business context. In the performance analysis, both intuitively and automatically selected significance levels have been used to perceptualise the impact of significance levels to scale semantic information and business context in both big data collection and analytics.

1.5 Structure Of This Thesis

This section provides an overview of the organisation of the remainder of the thesis.

Chapter 2: An overview of big data analytics - This chapter provides a survey of the status of relevant research areas: (i) applying contextual information in the applications needed for business organisations; (ii) business process management and the connections between business processes and business strategies; (iii) the development of semantic similarity in big data. Finally, it concludes with the potential research challenges identified in this chapter.

Chapter 3: A rule-based inference model for process-strategy relationship establishment - In this chapter an inference model is introduced. The aim of the inference model based on business rules is to establish the relationship between processes and the relevant strategies via their annotations and facts, respectively. By considering annotations and facts from the rules defined by business experts, automatic determination of the relationship can bridge the research gap associated with process-strategy relationship establishment. Allocating annotations to each business process reduces the handling complexity of a business process, while business strategy standardisation by using facts eases the ambiguity of business strategies.

Chapter 4: Significance level of big data text-based query - This chapter presents two approaches to the calculation of the significance level of a query. The first approach considers the weight of business processes to calculate the significance levels of queries. The second approach is an improvement to the first approach. This determines the significance levels of queries depending on the contributions of business processes and the priorities of the business strategies.

Chapter 5: Semantic information scaling and business context for big data analytics - This chapter presents a technique for embedding semantic information and business context into big data analytics. Based on the significance levels of queries calculated in Chapter 4, the use of semantic interpretation in both data collection and analytics is dynamically scaled. This technique is implemented in Hadoop using a query associated with a grocery shop using the automatically-calculated significance level as well as the manually selection of a set of significance levels for that query. Different sets of keywords representing the business context are used in this technique.

Chapter 6: Conclusions and future research - This chapter concludes the thesis and summarises the contributions of this thesis along with their potential benefits and impacts. Finally, the chapter highlights some research areas associated with the contributions of this thesis which could be explored further.

Chapter 2

An Overview of Big Data Analytics

Chapter 1 has highlighted the importance of context awareness and semantic manipulation in big data analytics. To enable a business application to provide real-time or near real-time big data analytics that can capture deep insight, it is of paramount importance to embed semantic information in both data collection and processing based on a business strategy. For this reason, this chapter presents an overview of big data analytic methods that specifically consider contextual information, the relationships between business processes and strategies, and the manipulation of semantic information. Based on the review of these techniques currently available in the literature, we identify a number of future research challenges that form the foundation of the research issues to be addressed in the subsequent chapters of this thesis.

The organisation of this chapter is as follows: Section 2.1 provides an overview of context awareness in big data analytics for business application. The role of business process and strategy in business organisations and the relationship between them are highlighted in Section 2.2. A brief review about semantic similarity for big data analytics is mentioned in Section 2.3. Section 2.4 describes the research challenges relevant to context awareness, business process and strategy, and semantic similarity in big data analytics. The chapter is finally concluded by Section 2.5.

2.1 Context Awareness in Big Data Analytics for Business Applications

Research on the integration of context in applications has been in existence for more than two decades. A number of applications exist which use context in diverse disciplines, including computer science and business, to improve business activities or system performance. This underpins the importance of the integration of context in business systems. Without context, the interpretation of a system, an event and even information may become meaningless. This is particularly true in data mining, especially in this information era, where without context, conclusions drawn from big data may be flawed [37].

Some techniques have been reported which consider contextual information in diverse areas (e.g., the context of a query and its users, the query-driven context-aware recommendations in [12,32], and a number of papers which review context-embedded methods in computer science, such as [3,42,43]. Some notable surveys of context-based methods and their brief discussion are presented in Table 2.1. In contrast, Table 2.2 presents a list of recent surveys on big data analytics applied to business applications, which shows that none of them have included context-based methods. However, since context can play a significant role in data validation and meaningful information extraction, it has been identified as a major vehicle for big data analytics [44]. The tables and our analysis of the literature reveal that there is another aspect in this field that has not been reviewed to date: context-based methods supporting enterprise applications. Given the recent attention to big data analytics in business applications, this review will provide researchers in this field with an overview of recent research in this area.

The aim of this section is to sketch a picture of context-based methods for big data analytics in the business domain [45]. Unlike reviews or surveys, which present comparisons of the methodological and technical aspects of context-based methods, this section provides an overview of what researchers have done, by considering context in big data analytics for applications supporting enterprise systems.

The structure of this section is as follows: Section 2.1.1 summarizes big data and the need for big data analytics in business organisations. The definitions of context, context models and context evaluation techniques for business applications are illustrated in Section 2.1.2. The main focus of this section is highlighted in four sub-sections: 2.1.3, 2.1.4, 2.1.5 and 2.1.6. These sub-sections reflect the most prominent aspects encountered in context aware applications for business organisations: context awareness in business process management (Section 2.1.3), context awareness in information retrieval (Section 2.1.4), context awareness in business intelligence (Section 2.1.5), and context awareness in big data tools and platforms for real- time data analytics (Section 2.1.6).

Table 2.1: Selected surveys of context-based methods

Title	Year	Description/Content
A survey of context-aware mobile computing research [46]	2000	A survey of the types of context used and the models of context information and systems that support collecting and disseminating context as well as applications that adapt to the changing context.
A survey of context-aware systems [47]	2004	Deals with context-aware middleware and frameworks.

A context modelling survey [48]	2004	A survey of the most relevant current approaches to modelling context for ubiquitous computing. It describes relevant approaches and classifies them relative to their core elements and evaluates them with respect to their appropriateness for ubiquitous computing.
A survey of context modelling and reasoning techniques [49]	2013	The requirements that context modelling and reasoning techniques should meet are discussed in this survey. These include the modelling of a variety of context information types and their relationships, high-level context abstractions describing real-world situations using context information facts, histories of context information, and uncertainty of context information.
A survey of context-aware workflow adaptations [9]	2008	Various existing approaches to adaptation in context-aware workflow are discussed.
A survey of context-awareness [50]	2011	The focus is context acquisition and sensing, context modelling and representation, context filtering and fusion, context storage and retrieval in context-aware computing. The development of context awareness and its applications is highlighted.

A survey of context data distribution for mobile ubiquitous systems [51]	2013	A unified architectural model and a new taxonomy for context data distribution, considering and comparing a large number of solutions, are presented.
A survey of context-aware recommender systems based on computational intelligence techniques [33]	2015	Surveys state-of-the-art context-aware recommender systems based on computational intelligence techniques.

Table 2.2: Selected surveys of business applications in big data analytics

Title	Year	Description/Content	Inclusion of context-based methods
Big data and management [52]	2014	This survey explores the potentials and opportunities for new theories and practices that big data might bring. It also presents several conceptual foundations, as well as possible avenues for future research and applications in business management.	No

Big data: A survey [4]	2014	A review of big data, related technologies (e.g., cloud computing, Internet of Things, data centres, Hadoop) and several representative applications of big data, including enterprise management, online social networks, medical applications, collective intelligence and smart grids.	No
Big data and its applications: A review [53]	2015	Several techniques involving big data which can be applied to various fields of engineering, industry and medical science are reviewed.	No
Challenges in big data application: A review [54]	2015	Big data management techniques and their challenges for business applications are discussed.	No

2.1.1 Big data analytics in business organisation systems

As stated in Chapter 1, data come from many sources and are hidden in many different forms [4]. This leads to business organisations being increasingly dependent on big data analytics. Hence, the development of new techniques that improve the effectiveness and efficiency of big data analytics is extremely important, especially for business areas where information technology directly/

indirectly affects businesses activities and their income [4, 55–57]. This section presents the impact of big data analytics on business organisations.

Recent studies show that the benefits of big data technology use in organisations are many-fold, including the production of more accurate results and cost savings [44, 55, 58, 59], as shown in Table 2.4. For example, an enterprise organisation can produce 61% more accurate results and save 56% of the cost [60]. Other substantial benefits are the retention and analysis of more data, increasing data analysis speed, and the reduction of manual processing [60]. The collection of valuable and relevant data and capturing their deep insights are not only a significant challenge for business organisations but also a major issue for the relevant research community.

Most businesses consider big data analytics as being mainly for revenue, customers and markets, as shown in Table 2.5. Based on big data, companies can innovate and enhance their current products and services, and develop and launch new products and services [52, 55, 58, 61]. Moreover, many advantages and benefits can be derived by applying big data analytics in different areas, such as customer intelligence, supply chain intelligence, performance, quality and risk management and fraud detection. In addition, industries such as manufacturing, retail, central government, healthcare, telecommunications and banking can also gain direct/indirect benefits from big data analytics [62].

Big data analytics play a key role in most economic areas, and this role is developing further, in parallel with the growing role of business intelligence applications in areas such as enterprise resource planning and marketing to leverage operational efficiency, competitive advantage and risk reduction. In addition, this development is highly relevant to the retail sector, where businesses can have more understanding about their customers. Therefore, they can improve their strategies for targeting consumers, and have faster fraud detection, and more accurate prediction [56]. There are certain areas where big data analytics add substantial value. An exploratory overview of relevant retail levers associated with big data is presented in Table 2.3.

Table 2.3: Big data levers in the retail industry (www.mckinsey.com)

Function	Big Data Levers
Marketing	Cross-selling.
	Location-based marketing.
	In-store behaviour analysis.
	Customer micro-segmentation.
	Sentiment analysis.
	Enhancing the multi-channel consumer experience.
Merchandising	Assortment optimization.
	Pricing optimization.
	Placement and design optimization.
Operations	Performance transparency.
	Labour input optimization.
Supply chain	Inventory management.
	Distribution and logistics optimization.
	Informing supplier negotiations.
New business models	Price comparison services.
	Web-based markets.

Table 2.4: Benefits of big data technologies in enterprises

Retain and analyse more data	74%

Increase speed of analysis	70%
Produce more accurate results	61%
Reduce or eliminate manual processes	59%
Cost savings	56%

Table 2.5: Top 5 priorities companies hope to gain from data analytics/big data investments

Revenue	62.5%
Better understanding of clients/stakeholders	57.3%
Better understanding of markets/marketing analysis	57.1%
Productivity gains	54.3%
Customer acquisition	51.1%

Over the years, the process of making managerial decisions has become of paramount importance, and it has been a challenging research topic. In the view of most decision-makers, the value of big data depends on the ability to deliver key information and knowledge to enable better decisions. Since big data are generated from a variety of sources with different representations and formats, the question for any researcher or business is how important data relevant to a business context can be captured and analysed more accurately to represent deep and relevant business insights [52,55,58,59,63].

Therefore, the analysis of big data has become a very important asset for making decisions, and offers the opportunity to deliver significant advantages and benefits to business organisations [55, 61, 64]. However, this will come about only when data are captured and analysed appropriately and effectively to uncover the deep insights of the data. Better decisions can be made upon the subsequent changes from historical and real-time data generated through means like supply chains, customer behaviours and production processes [44,52,55].

According to Elgendy and Elragal [62], business organisations not only focus on analysing internal data (e.g., sales, customer feedback via phone or email, shipments and inventory) but they also need to consider data from outside sources such as social media and forums. "With the increasing sizes and types of unstructured data on hand, it becomes necessary to make more informed decisions based on drawing meaningful inferences from the data" [62].

2.1.2 Context definition, models and evaluation for business applications

2.1.2.1 Context definition

A number of recent research studies have shown that the use of context in interactive applications is very useful and important [65–69]. By considering context in computer applications, it has been demonstrated that the number of irrelevant results is reduced and more accurate data are collected. For example, the application of a query context and user context in information retrieval improve retrieval precision substantially [8,12]. In addition, the term "context" has been embedded in information technology, business and other application domains for a long time, and has improved the efficiency and effectiveness of these applications [12,70].

Due to the richness of everyday language and the knowledge that people acquire during their life-time, humans find it easy to communicate and understand each other and they also react appropriately, based on the context surrounding them. However, a machine like a computer does not have the ability to convey ideas, transfer information well to humans, or interact with them [71]. Therefore, the richness of communication in human-computer interaction will be increased by improving the computers access to context, and as a result it is possible to produce more useful computational services [49].

To fulfil the purpose of using context successfully, people must understand what context needs to be integrated, how it can be used, and when context should be considered [9, 49, 71]. In [43], the authors emphasize that an understanding of context would enable application designers to choose what context needs to be used in their applications, and provide them with an understanding of how context could be used. This will help application designers determine what context-aware behaviours to support in their applications.

The terms "context" or "context definition" first attracted the attention of the computer science community in the 1990s. After their introduction in 1994 [72], there was an explosion of definitions from 1997 to 1998. Table 2.6 shows some of the first generation of context definitions that have been most popular and most widely used.

Table 2.6: Early definitions of context

Author	Definition of context
Schilit & Theimer (1994) [72]	Location, identities of nearby people and objects and the changes to those objects.
Brown et al. (1996) [73]	Location, identities of the people around the user, the time of day, season and temperature.

Ryan et al. (1997) [74]	A users location, environment, identity and time.
Dey (1998) [75]	A users emotional state, focus of attention, location and orientation, date and time, objects and people in the users environment.

However, Abowd et al. [43] argue that the context definitions presented in Table 2.6 are not easy to apply in any system or application. For example, when one wants to determine whether a type of information not listed in the definition is context or not, it is not clear how to use the definition to solve the dilemma.

Therefore, in 2001 another definition, which has become the most popular, was proposed by Dey: "Context is any information that can be used to characterize the situation of an entity" [71]. An entity can be a person and place or any object which is considered relevant to the interaction between a user and an application.

In the same year, Winograd [76] presented a more specific definition as follows: "Context is an operational term: something is context because of the way it is used in interpretation, not due to its inherent properties". Coutaz et al. [77] defined context in 2005 as " not simply the state of a predefined environment with a fixed set of interaction resources. It is part of a process of interacting with an ever-changing environment composed of reconfigurable, migratory, distributed, and multi-scale resources".

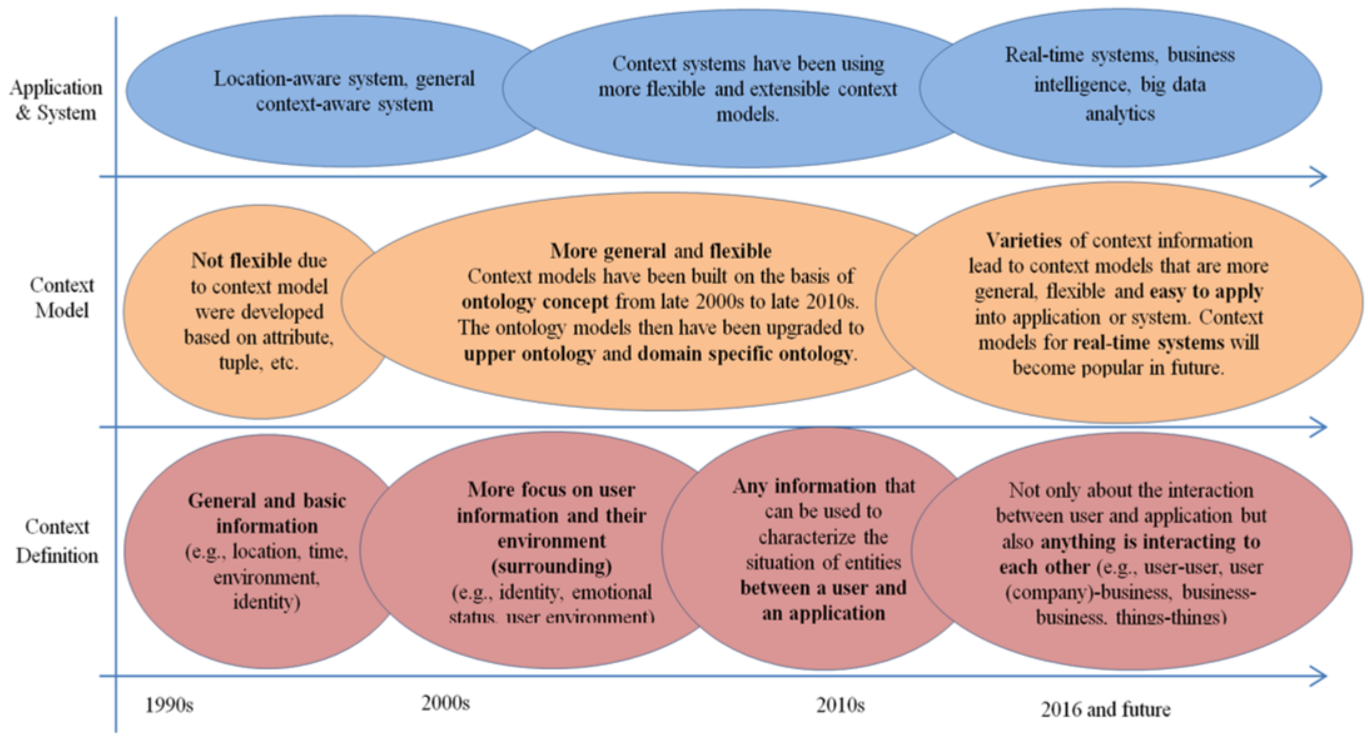

Figure 2.1: The development of context definitions and their applications

Rapid technological development and the advent of different data sources including big data have produced a large number of diverse applications and systems at an increasingly higher rate. This has driven researchers and application developers to change the representation of context over time. Figure 2.1 shows that context initially meant simply a location.

However, over time it has been extended to represent any information about the situation of an entity. In the 1990s context-aware applications and systems were limited to location-aware systems, but later in the 2010s they were extended to any system which uses a more flexible and extendable context model.

As mentioned above, many definitions of context have been introduced. However, the embedding of a context into an application/system requires its modelling. Based on the definitions of context, various context models have been developed. The next section presents some contemporary and widely-used models of context.

The six phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM) are: (i) business understanding, (ii) data understanding, (iii) data preparation, (iv) modeling, (v) evaluation and (vi) deployment [1]. The early phases of the CRISP-DM model, namely data understanding and data preparation, must consider data characteristics for mining purposes, and therefore these phases, when embedded within business contexts, have direct impacts on big data analytics. Context-aware methods have been found to be effective when contextual information is considered in analytics. The following Table 2.7 describes the data understanding and preparation stages of CRISP-DM, considering big data models and contextual information.

Table 2.7: Big data models for data understanding and preparation of phases of CRISP-DM in relation to context awareness

Dimension of big data	**Phases of the CRISP-DM model: Business/ Data understanding**	**Phases of the CRISP-DM model: Data preparation**
Variety	Since the context of a business organisation is developed considering its general business strategies, context can be applied in gathering different types of data from various sources that are essential to pursue the business strategies.	Eliminating irrelevant data using a context during the data collection phase will eventually reduce the load of the data preparation stage. This concomitantly enables the data preparation stage to process and transform computation- and storage-intensive qualitative data (e.g., text, audio and video) into a form (e.g., key-value pairs) suitable to be handled by big data modelling tools like Riak, Oracle NoSQL, Cassandra, DyanmoDB, MongoDB, and Azure Table Storage.

Continued on next page

Table 2.7 – continued from previous page

Dimension of big data	Phases of the CRISP-DM model: Business/ Data understanding	Phases of the CRISP-DM model: Data preparation
Velocity/ Volume	Continuous generation of data from various sources at high velocity creates a huge volume of data which is difficult to process in real-time and store for future processing with existing technologies. Sampling of these data can address the challenges related to both volume and velocity. An up-to-date business context can be used to fine tune the sampling policy in order to tackle these challenges.	Data preparation is a time-consuming phase of data mining [1]. Since the context of an organisation is mainly derived from its business goals, it is highly likely that the types of data and their sources being selected through its context will be the most relevant to that organisation and hence meet the organisational plan. Other studies [78] show a significant reduction of volume and increase in relevancy in data collection.

Continued on next page

Table 2.7 – continued from previous page

Dimension of big data	**Phases of the CRISP-DM model: Business/ Data understanding**	**Phases of the CRISP-DM model: Data preparation**
Veracity	Quality data is a precondition for acceptable analytical outcomes. In contrast, poor data quality due to problems of accessing reliable and trustworthy data from existing sources poses substantial hindrance to the improvement of business strategies.	Since big data are not usually structured, all data are not clean and precise. Contextual information can be utilized to identify the inaccurate and missing data in the data cleansing task in this phase. One study [78] suggests this eventually improves the reliability and trustworthiness of big data.

Continued on next page

Table 2.7 – continued from previous page

Dimension of big data	**Phases of the CRISP-DM model: Business/ Data understanding**	**Phases of the CRISP-DM model: Data preparation**
Validity	As context reflects the data requirements for a business plan, it is highly likely that data collection through a context will be more relevant to what a business wants to achieve and more useful for its intended usage.	The context of a business organisation is developed considering the information relevant to a particular application. Therefore, data validation, one of the key recommendations for data construction, using a business context increases the likelihood of meeting all application-specific criteria for the intended use of the data [78].
Volatility	To indicate the duration for which data remain valid, the data expiry period can be added to the context description, and then it can be used to avoid collection of expired data.	A data expiry period, one of the pieces of contextual information, can be used to clean outdated data automatically from big data storage.

Continued on next page

Table 2.7 – continued from previous page

Dimension of big data	Phases of the CRISP-DM model: Business/ Data understanding	Phases of the CRISP-DM model: Data preparation
Value	As indicated above, for data understanding, context can be exploited to address the issues related to the other 6Vs (Variety, Velocity, Volume, Veracity, Validity and Volatility), and by using contextual information, the context-aware methods are capable of collecting valuable information towards fulfilling an organisations business objectives. All these will eventually increase the chance of achieving more value from the collected data.	Similar to data understanding, since contextual information can be applied in solving a number of matters of other Vs in data preparation, it will increase the value of the data considerably.

2.1.2.2 Context models

Context definitions include only the abstract level of information for representational purposes. However, to embed content in a particular system, the context has to be designed (modelled), taking into account all its essential structural fine details so that it can meet the purpose of that system. Therefore, a context model is a set of information coming from external and/or internal systems that can have some effect on its relevant system [49]. As shown in Figure 2.1, original context models in the early 1990s were less flexible, as they considered only the attributes or value tuples of location.

However, models have become more flexible and extensive since 2010 [34]. Depending on the type and the requirements of the systems, many different context models have been developed. Context models that have been applied in the business domain can be classified into three main groups: i) the n-gram model, ii) the tree structure model and iii) the onion model.

An n-gram context model is mainly used to present a query context in information and document retrieval [12]. The most popular tree structure models are: i) user context, ii) B2B (Business-to-Business) context models, iii) business context models for big data collection and processing and iv) context trees for a business processes. In [3], a user context model for an information retrieval system assimilates all the factors that describe a users intentions and perceptions about the surrounding factors, as shown in Figure 2.2.

In this model, the user context is divided into five main sub-contexts i) task, ii) social, iii) personal, iv) spacio-temporal and v) environment. The task context includes goals and task information, while the social context is information on social contacts such as friends and colleagues. Personal context is represented in terms of physiological (e.g., height, weight, age) and mental (e.g., interests, domain, expertise) contexts. The social-temporal context refers to location, direction and time, while the environmental context comprises light, services and people [12].

As an outcome of a number of research studies on B2B collaboration, the B2B context model shown in Figure 2.3 has been proposed to support supply

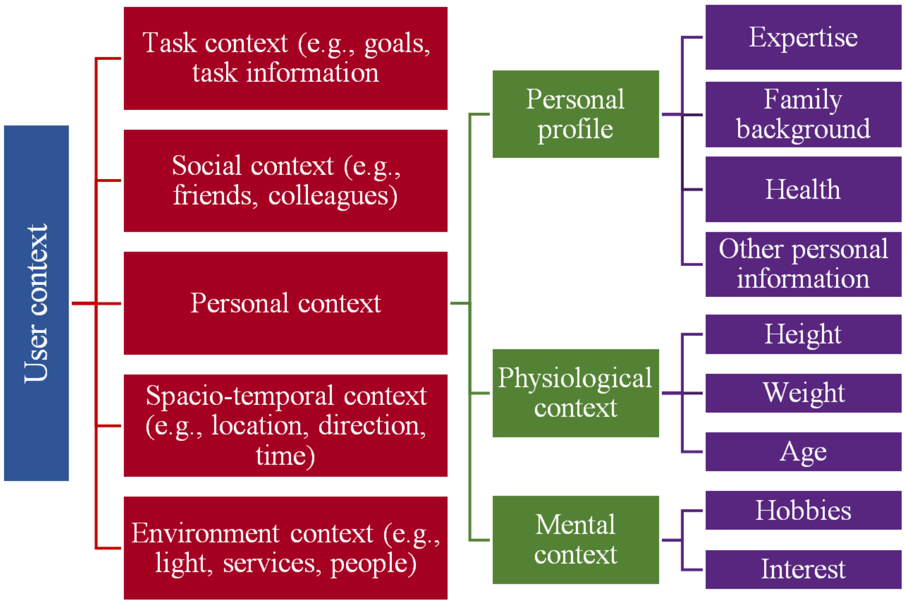

Figure 2.2: An example of a user context model

chain applications [79, 80]. In this model, information on the user or company (e.g. user expression, profile, industry sector, product service, track record), temporal (e.g., time expression, periodicity, lead time) and location (e.g. geographical expression, transportation mode, political stability index, tariffs) is used. Since this model was developed particularly for B2B collaboration, the main focus is not on the business system processes. For this reason, it is not effective for business process modelling, and this paves the way for the development of a context model considering the main aspects of a business process.

A tree structure context model based on the basic taxonomy of a business context which captures most common context-related knowledge has been developed by Saidani et al. [81]. This model consists of essential aspects of a business process, including resources and organisational units. For example, resources include business objects and actors and organisational units (e.g., quality of communication, quality of relationships, actors proximity) that belong to organisations in the context tree.

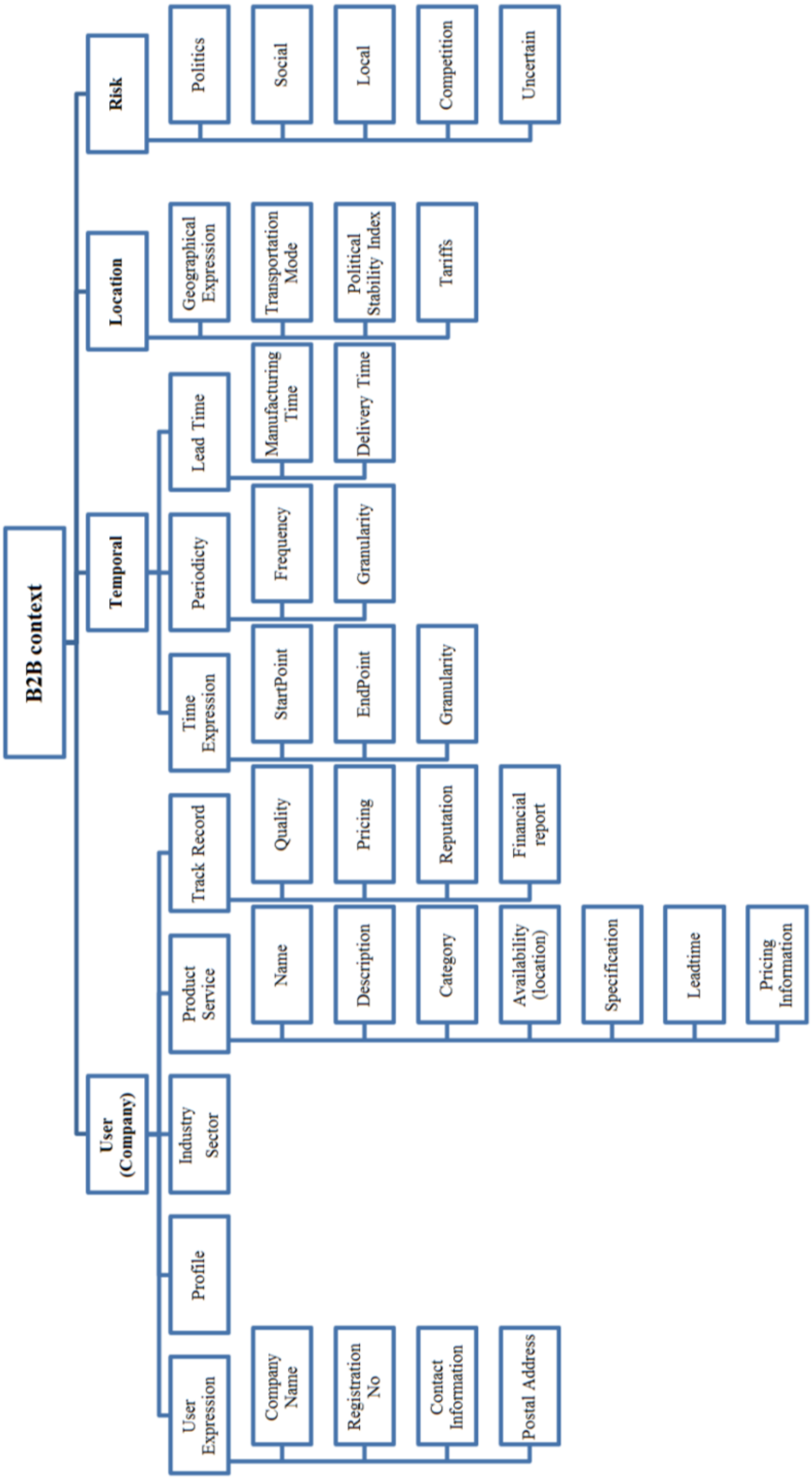

Figure 2.3: Part of B2B context model

Rapid economic and technological progress makes the expectations of customers diverse and ever-changing, resulting in changes in the context in which expectations are formulated. Therefore, there is a need to consider context-related knowledge in the behaviour of a business process [81]. As customers are the key output value, a context model for a business process must consider the business organisation and human beings, such as user context, environmental context, and the internal context, including organisational and goal contexts [82]. Considering the contextual information, a model has been proposed [82] which is called the onion context model. This is shown in Figure 2.4.

A large number of methods/applications considering contexts have been reported in the literature. However, as mentioned in 1, the main focus of this chapter is to review methods/applications that use contexts for processing and analysing big data for business organisations. In the remaining sections of the chapter, we consider some of the most common context-based methods that have been used in businesses, such as business process management, information retrieval and business intelligence.

Since many different types of contextual models exist in the literature, this raises an important question: what types of contextual information are essential for context-based applications in the business domain and how can we ensure that they are appropriate for a business? The former requires context modelling techniques, while the latter necessitates context evaluation. The next section addresses this issue.

2.1.2.3 Evaluation of contextual information

Generally, context evaluation can be performed in two ways: evaluation of (i) a context model and (ii) a contextualised business process (i.e., a business pro-

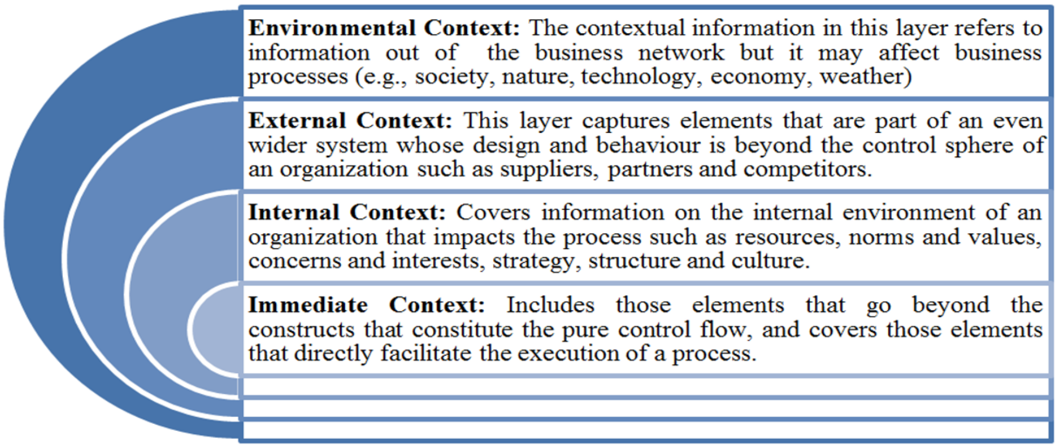

Figure 2.4: Onion model for context classification

cess which embeds context in its bedrock). They are described in the following sections.

(i) Evaluation of a context model

Thus far, we present various types of context model structures used in many different applications in Section 2.1.2.2, but how they were developed has not been discussed. Before evaluating and determining the efficacy of context models for business applications, it is important to consider how a context model is usually developed. In developing context-based methods, the first and most important stage is to model contextual information. To do this, the basic and most popular way is to build a model based on observing and examining the attributes of a business system and its users and physical environments. For example, what types of contextual information a business system needs, what contextual information including environmental attributes may influence its users, what information a user may need at the time they work on the system, and what information a business system needs to be aware of when it is going through some changes. Is there any relationship among these pieces of information?

As shown in Figure 2.5, there are five main steps in context modelling: (i) learning what information a system needs, based on observing the activities of the system and the behaviours of users, (ii) making a prediction of what contextual information is potentially required, (iii) deciding on what information needs to be considered in the context model, (iv) indicating the relationship among the types/pieces of contextual information based on practice and experience, (v) building the relationship among contextual information, users and systems [83]. Although these steps have been mainly applied to context modelling in information retrieval, they have also been used in other context-based areas, such as business process, business intelligence and real-time analytics. By following these steps, more accurate context models can be developed [83]. Although a context model can be constructed following the steps presented in Figure 2.5, the important question remains: how can this context model be evaluated appropriately in business applications? To address this issue, the model of a context can be divided into a number of fundamental contextual components. For example, the user context shown in Figure 3 has been divided into five contextual components: (i) user expression, (ii) profile, (iii) industry sector, (iv) product service and (v) track record. For each of these components, similar to the unit or module test applying the white box testing technique in software engineering, some suitable test cases can be derived considering all possible, important and relevant contextual scenarios to ensure its validity for all aspects and assumptions, and its effectiveness and appropriate adaptation in a dynamic business process. Similarly, after testing and modifying all individual components, they can be combined. Some test cases can be derived for integration testing to verify whether the components work together [83]. The generic nature of a context model can be tested by applying it in different types of applications. Therefore, to test whether a context model (e.g., tree context model, n-gram context model, contextual graphs) is generic, we need to test whether it is applicable in other types of applications with minor modification or extension [84]. Another way to evaluate a context model is to assess a context model and its components by applying the contextual analysis technique introduced in [10] for a particular contextual scenario. This is detailed in the following section.

(ii) Evaluation of contextualised business processes In a business process, contexts are analysed and evaluated based on determining the context variants via truths (e.g., the facts that shoppers asked about a promotional campaign, that a public holiday was in the region of the store) and statements (e.g., a shopper showed interest in a promotion or lack of it for being in a hurry). A process contributor (e.g., process designer, marketer, or salesman) is able to verify a truth. However, they are unable to verify a statement. The use of context analysis enables reasoning about the context in the business process, as well as the discovery of other contextual information that may relate to the current context.

Furthermore, contextual analysis based on the truth of context properties and the statement of assuming truth value can provide a deeper view for a particular context. If a context model is developed using a particular technique (e.g., using the steps shown in Figure 2.5), the validity of such a model for all possible business scenarios can be examined using the context analysis technique introduced for a contextualised business process [10].

For example, a sample context analysis technique for the scenario, "finding a potential shopper", is described in Figure 2.6. In this scenario, the truth (T refers to ¬(T1 OR T2) AND T3 where T1, T2 and T3 are defined in the right bottom corner of Figure 2.6) indicates that although a shopper is neither asking for the product (P) which is on promotion (Pr) nor requesting information about Pr, s/he may be a potential shopper (S1) if the region in which the store is located is on holiday. After finding a potential shopper, if it is found that the shopper is not interested for any reason (e.g., the shopper does not want to be disturbed when s/he is in a rush), the process model should revert to the context analysis technique for finding potential shoppers. Therefore, the business

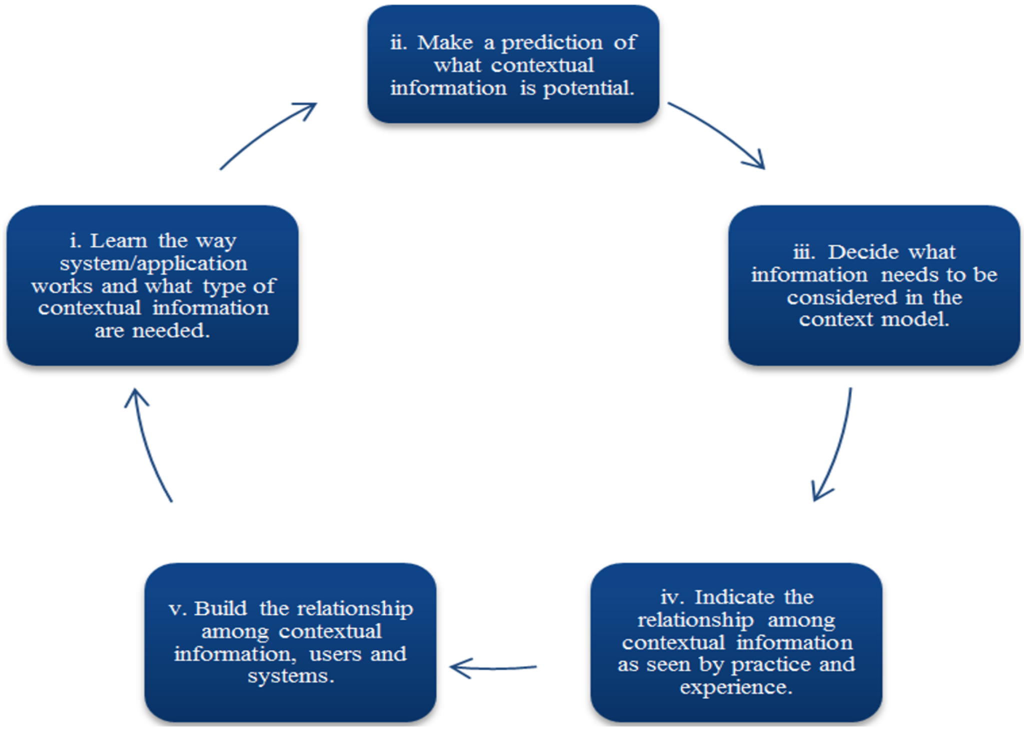

Figure 2.5: Circle of development of context model

process needs to be updated with the purpose of improving its adaptability to context variants and contextual changes. As a result, a new contextualized business process which is more effective and adaptable to context changes for the sale promotion program can be developed and executed.

Contextualisation of a business process via context analysis can guarantee to output a better business process model. In addition, the incorporation of context analysis into a business process can validate that a contextualised business process fits its context and is effective [85]. During the implementation, using a technique used in software engineering, namely black box testing, the functionality, behaviour and reliability of a contextualized business process can be tested by matching the actual outputs produced by the system with their respective expected outputs under various contextual scenarios. The black box testing technique mainly considers test strategies, such as customer requirements, equivalent partitioning, boundary value analysis, decision table testing and test case failures [86].

Finally, the accuracy and coverage of information in context models are high, except for the context models applied in big data analytics including real-time analytics, which is caused by technical limitations. In addition, the adaptability in their applications may range from low to high, depending on the complexity and rapidity of context changes in the area where they are applied, as described in Table 2.8.

Table 2.8: Subjective evaluation of contextualized systems for various types of business applications

Area	**Research trend**	**Type of contextual information**	**Accuracy and coverage of information**	**Adaptability of context model in application**
Business process workflow	Early stage	People (e.g., customer, staff), activity/ process (e.g., buying or selling products, interacting with customers, returning or refunding orders), organisation (e.g., goal, strategy, organisational structure, functional areas, products), environmental or external factors (e.g., competitors, suppliers, economy, location)	High accuracy but low coverage (e.g., lack of contextual information for competitiveness, customer interest) [86].	Adaptability of context models in this area is low due to the complexity of contexts, the changing nature of contexts and the variety of business activities (e.g., competitiveness, weather, customer interest and resources).

Information retrieval, recommendation systems, query classifications for web search and online advertisement	Established growth	User and environment	High accuracy and medium coverage (because which context information of a video or image should be included in the context model to make applications in this field work more effectively remains a major question)	Medium adaptability due to established research and low application requirements, context models in this area generally meet the needs of their applications. However, their adaptability is medium, as capturing context information of some objects (e.g., image, video) is still challenging.

Business intelligence	Early growth	People (e.g., customers, staff), organisation (e.g., goal, strategy, organisational structure, functional area, products), environmental/ external factors (e.g., competitors, suppliers, economy, location)	High accuracy but low coverage (e.g., lack of contextual information for competitiveness, customer interests and real-time analytics). Context information for business intelligence should incorporate business indicators and predictors [87].	High adaptability. Business intelligence is currently focused on analysing internal data. As a result, context information for this type of application is not changing rapidly. Therefore, the adaptability of context models is high. However, the adaptability may become very low in near future because of context changes and making real-time decisions. Due to the development of social media, business intelligence should also consider external information such as customers comments and opinions on social media and product consumption trends to help businesses make better decisions.

Big data tools and platforms, real-time analytics	Early growth	User and environment	Low accuracy (due to technical limitations and difficulty in capturing information in real time). Low coverage (contexts change rapidly in real-time analytics).	Low adaptability. Context changes, volume of big data, latency of streaming data and technical limitations make the adaptability of context models for context-aware applications in this area still low.

Table 2.9: A comparison of business process and workflow

Comparative aspects	**Business process**	**Workflow**
Definition	A business process is a collection of activities that takes one or more kinds of input and creates an output, such as a report or forecast that is of value to the customer. [2]	The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules [88]
Similarity	They both foster mainly a process-oriented perspective on organisations [89]	
Differences	Includes manual activities/processes	Focusses on processing of digital business/office documents
		Continued on next page

Table 2.9 – continued from previous page

Comparative aspects	Business process	Workflow
	"Conceptual level" of the enterprise	Manual processes or decision-making processes are not considered
	Business processes can be related to every kind of resource i.e., are not limited to information technology.	Emphasizes use of information technology. Workflows are business processes combined with information technology

2.1.3 Context awareness in business process management

In information systems research, business process (BP) and workflow (WF) are two key foci to improve the efficiency and productivity of a business organisation [89]. Although business process and workflow sound very similar, they concern the same matter from two different points of view [89]. An organisational process-oriented perspective comprises activities and their relationships within and to an organisations context required for both business process and workflow [89]. Table 2.9 shows their definitions, main similarities and differences.

There are many ways to improve effectiveness in managing these areas [88, 90, 91]. However, given the scope of this chapter, only approaches that consider contextual information are presented here. Whichever business process or workflow is applied, the use of contextual information in these areas

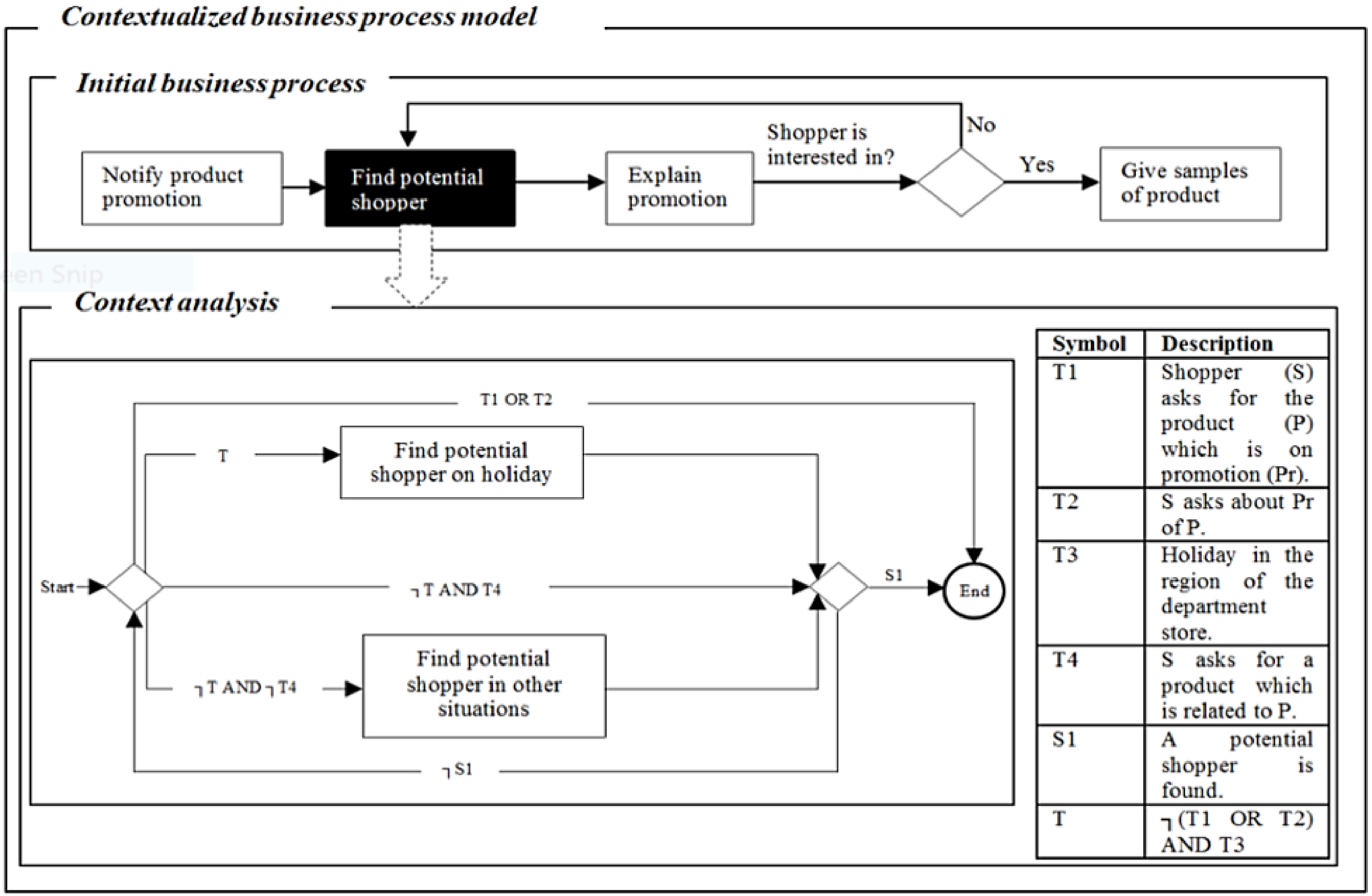

Figure 2.6: Verification of contextual information for contextualised business process model for a particular business process scenario, namely find a potential shopper

has the same purpose: improvement of the capability of systems by recognizing and reasoning context during the life of a business process or workflow. As mentioned above, the output of a business process must be of value to customers [2]. The customer may be a traditional external customer (e.g., one who buys a product) or an internal customer, such as a staff member of the same organisation.

An effective workflow management system is one of the strengths that helps businesses achieve goals by sequencing work activities and providing the appropriate human or information resources associated with these activities [92]. Many different kinds of approaches are available in the current literature to improve effectiveness in workflow management systems, one being context-aware workflow systems [9, 11]. According to Smanchat et al. [9], "to implement workflow in a context-aware system, the important features of context-aware systems which are context-awareness and adaptability must be accommodated into the workflow".

While business process management (BPM) has been shown to help organisations improve and innovate, Brocke et al. [93] claim that one essential problem related to this development is that the BPM body of knowledge does not account for a broader variety of business contexts. The authors also stress that BPM needs to be contextual for the purpose of increasing efficiency and effectiveness. BPM has been recognized by a number of researchers and will become more popular in the near future. Most recent approaches focus on how to model context for business process and workflow, or examine which contextual information is suitable for business processes.

Based on our review, we conclude that consideration of contextual information in business processes/workflows is still in its infancy [56,86,94]. Some approaches that have been introduced for business processes are shown in Table 2.10, which presents the contextual information and modelling approach for each of them. The main issues that all relevant research communities want to solve are how to: (i) identify appropriate contexts for business processes, (ii) embed them into the processing of business activities effectively and efficiently and (iii) derive and update the context model dynamically from the information collected.

Table 2.10: Context-based methods in business processes

Author	**Year**	**Contextual Information**	**Approach/ Description**
Chen et al. [95]	2001	Internal information, external partners, systems and resources	Relevant BPs across participating organisations are integrated for efficient functioning of business in the global market (collaborative business process).

Roseman et al. [82]	2008	The onion context model shown in Figure 2.7 is presented in four layers: environmental context such as society, nature, technology, economy and weather; external context such as suppliers, partners and competitors; internal context such as resources, norms, values, concerns, interests, strategy, structure and culture, immediate context such as those elements that go beyond the constructs that constitute the pure control flow	Based on a goal-oriented process modelling approach [96]; a given process and the appropriate measures are used to reason about the potential contextual elements that are relevant to the process and thus these elements are to be included in the process model. Context elements are then classified into four different types of context layers based on the onion context model.
J.L. de la Vara et al. [10]	2010	Context is found based on a goal-based framework for contextual requirements modelling and analysis [85].	Analysing business process context and then modelling the business process after business process contextualization based on a goal-based framework for contextual requirement modelling and analysis.

Multiple parties work in a collaborative manner in intra-enterprise and inter-enterprise process management. Parties are unlikely to work with a centralized process management because of the preservation of self-interests, and the security and privacy of shared data. Rather, this requires peer-to-peer interactions between the parties. Such collaboration among multiple parties is emphasized by Chen et al. [95], as "relevant BPs across participating organisations are integrated for efficient functioning of business in the global market and called collaborative business processes". Therefore, BPM considers not only internal information but also external information exchanged among partners. Tan et al. [65, 79, 80] visualize B2B collaborations aided by a context-aware framework which categorizes contextual information into user (i.e. the company), temporal and location. Most B2B transactions are dominated by the exchange of business documents, which are rich in contextual information in the form of user, temporal and location information. However, the main criticism of context-aware B2B approaches is that the future of context awareness very much depends on a theoretical breakthrough in the underlying enabling infrastructure and not simply extensions of current architectures. History has shown that advances have been marked by progress to what is convenient and natural to use (especially in computer science). In addition, B2B information exchange standards do not embody contextual information, whereas the integration of contextual information into web services supporting the underlying processes is regarded as the first phase of a context-aware business process system [97].

Rosemann et al. [82] and Silva et al. [98] argue that most existing approaches concentrate on intrinsic ways of adding or modifying business processes after the need for process change arises. They emphasize that the actual drivers of flexibility have not yet been discussed thoroughly and as a consequence, current process modelling techniques capture only the reactive part of process flexibility, but lack contextualization [82]. For this reason, they introduce BP systems considering the onion context model shown in Figure 2.4. This model classifies the contextual information into four different context layers: (i) environment, (ii) external, (iii) internal and (iv) immediate. The application of these context layers makes these systems more realistic and responsive to the relevant external triggers [77]. Although BP systems based on the onion context model cover various types of contexts of a business organisation and the relevant entities among which an organisation mainly communicates, they are adaptable to the dynamic variation of contextual requirements for a business process model instance [99]. This is because there is a significant difference between context and semantics/ontological matching, and this raises the question whether the ontologies of different companies can ever be matched [100]. This has motivated researchers to develop a business process context via context analysis (e.g., SWOT analysis). Inherently, this increases the likelihood of having the semantics/ontological matching of a context with its business organisation and helps determine the essential information that impacts the design and implementation of a business process [44, 92]. All these research studies gave rise to the development of two structures for modelling and categorizing the context: (i) the context tree depicting contextual characteristics; and (ii) an adapted context tree applicable to a specific domain [85].

When a context is analysed during the modelling of a business process, the identification of all its variants (relevant states of the world in which the business process is executed) and the determination of how the business process should be executed in them occur simultaneously. Therefore, the modelling of both a business process and its context should be conducted together and optimised using an iterative process. This indicates that research on this topic is still at an initial stage and many challenges have not been addressed to date [10, 101, 102]. An approach has been introduced by Vara et al. [10] which considers a business process and its context analysis through an iterative process. This approach includes four steps: (i) an initial version of the business process that needs to fit its context is modelled, (ii) relevant context variations influence the business process and imply that the business process execution has to change, (iii) if a context variation is found, the business process context is analysed to find the context properties that allow process participants to know if a context variant holds. A context analysis model is then created, and context variants of the business process are analysed, (iv) a contextualised business process model is created on the basis of the final context variants and their effect on the business process [10]. This approach loops through steps (ii) to (iv) until context variation is not found [10].

As the consideration of context and its relation to big data analysis has been the focus of the data science community, a number of existing data analytic methods consider contextual information for information retrieval. These are presented in the following section.

2.1.4 Context awareness in information retrieval

Information retrieval (IR) is used to collect the required information from various data sources (e.g., database, web, big data) through a query. It refers to both text document, and content-based image, audio and video retrievals. IR has a wide range of business applications, including data mining, search engines, robotic vision, e-commerce, entertainment, opinion formation, automatic matching and rating. Context-aware IR has been in existence for many years. The concept of context has been used in IR to develop a context-based video retrieval system which utilizes various sensor data to provide video browsing and retrieval functions for life-log applications [103], or integrate user context and query context in an information retrieval model [12]. Unlike data mining or business intelligence, IR focuses on how to collect the information which users need without capturing deep insights from the data [104,105].

Although an information retrieval system sounds easy to develop since it concentrates on raw data collection, due to big data, the variety of data types as well as the volume of data size keep increasing significantly. Therefore, the collection of information on what users need has become a difficult mission for any system. A study conducted by Aknouche et al. [12] on the perception gap among collecting relevant information emphasizes that most information retrieval systems rely on the retrieval decisions made by queries. As a result, these systems yield information related to actual users. Since the search context is ignored, a large number of irrelevant results occur in these systems [12].

In [8, 12, 32, 33, 70, 106], IR techniques are proposed which integrate both user context and query context based on language modelling. The query context includes the integration of linguistic and semantic knowledge about the user query (e.g., instead of explicitly assigning meaning to a word, the differences between word usages are made based on different contexts and context words are added to relations in order to exploit another relevant word context within the query). User context can be found either by the users domain of interest or the topic of interest. A part of a sample user context is shown in Figure 2.2. For example, if a user mentions "Apple" in a query, then there is difficulty in recognizing whether he/she is interested in fruit or the brand name of a technology company. Many existing methodologies classify the query into both categories of "Fruit" and "Technology Company" without understanding the user's search intent. However, by considering the users domain of interest or topic of interest, if the query "Fresh" is found before "Apple", the category of "Fruit" will be assigned to the user interest. In contrast, if the user issues some queries related to technological items such as laptop and computer before "Apple", the user interest will be "Technology Company". User context plays a vital role in information retrieval and to date no information retrieval system takes into account all intentions and perceptions of users.

Query classification uses the search context of a query to perceive the purpose of that query. This covers many important business applications, such as online advertisement and big data search. Unlike traditional methods that are non-context-based and focus only on individual queries lacking exact context, in recent times, query in information retrieval has shifted its focus to context awareness. Examples of work emphasizing context in query-based retrieval include a context-based query classification considering neighbouring queries and user actions [70, 107] and a query classification method using conditional random field models taking context into account [32,70]. In general, all current information retrieval and data mining methods face great difficulties because they are expected to be able to handle many challenges that big data face [108]. The challenges include, amongst others, unprecedented heterogeneity, speed, accuracy, volume, privacy and trust; however, these challenges are yet to be met.

According to [3, 4, 55, 57, 58] a large number of research projects have been undertaken in order to improve the existing methods and techniques to overcome the challenges in big data. Methods proposed include the application of massive parallel processing architectures and novel distributed storage systems, metadata repositories of the contextual information of a business, and innovative mining techniques based on new frameworks/platforms (e.g., Hadoop, Mahout, Giraph). These techniques have the potential to successfully overcome the aforementioned challenges and reshape the future of data mining technology [55].

Recommendation systems for information retrieval provide tailored information on a user, service or product by filtering the query results. Similar to IR, recommendation systems have many business applications which include, but are not limited to, entertainment, e-commerce (e.g., online shopping, marketing, and advertisement), opinion formation and the rating of a particular service or product. Research shows that context-aware recommendation systems produce more appropriate recommendations than their traditional counterparts such as non-context-based systems [95]. An example of such a system is a query-driven context-aware recommendation. This is an extension of the Latent Dirichlet Allocation (LDA) [32], where the combination of users, items and the meta-data associated with contexts are considered, as shown in Figure 2.7.

The model is shown in Figure 2.7 and integrates user profile, item descriptions and contextual information. The authors of [32] state that a context-aware system may hold contextual information which is relevant to a situation or a task in a variety of ways. For example, a user of a music recommender system can specify his/her current interest in a specific genre of music by providing that information in the form of a query [32]. While contextual information is becoming important in business process and information retrieval, its application is not common in the business intelligence (BI) domain. However, contextual information can benefit BI and has already been exploited by a few approaches for BI, as described in the following section.

2.1.5 Context-based methods for business intelligence (BI)

BI plays the main role in decision making for any organisation, including business systems. It consists of many processes and technologies. These help an enterprise transform raw data into meaningful and valuable insights or clues that can contribute to identifying and improving inefficient business processes. As a consequence, this enhances the traditional decision support system (DSS). BI systems are not only useful for businesses but also for government organisations and other sectors such as banks, agriculture and health care. A number of research projects on BI and its different frameworks and tools have been conducted. The application of BI in different enterprises is distinct in nature. However, the most contemporary conceptual framework of

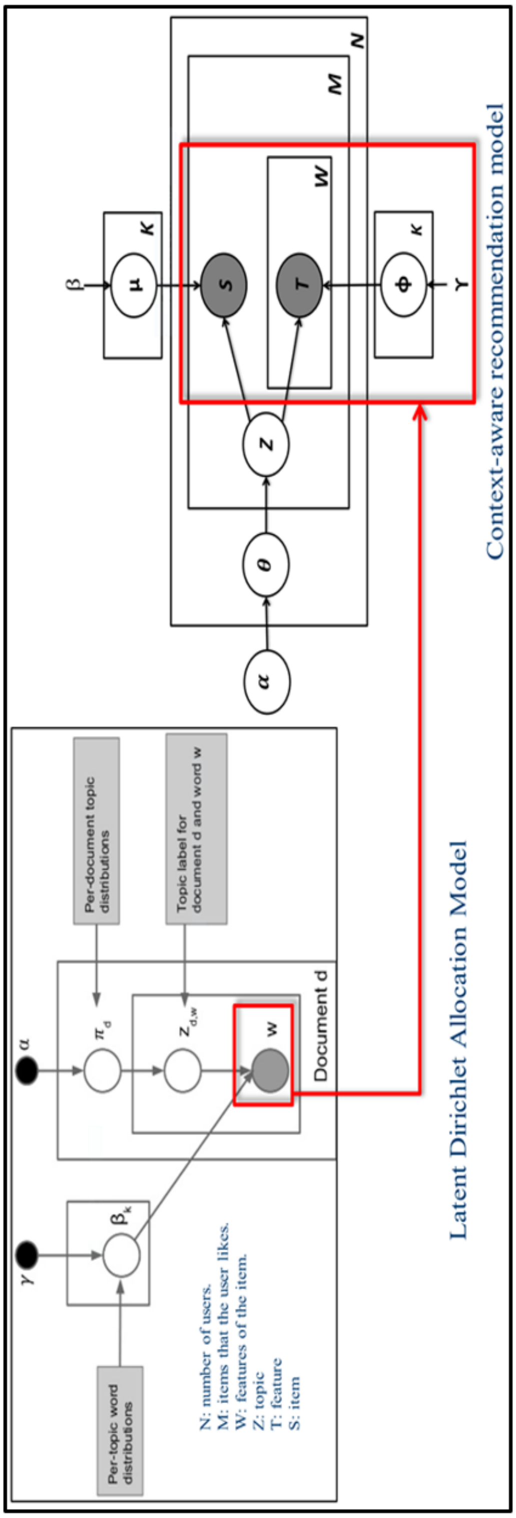

Figure 2.7: A context-aware recommender system

BI can be represented by the diagram shown in Figure 2.8. This comprises five main phases: (i) data sourcing, (ii) data analysis, (iii) situation awareness, (iv) risk assessment, and (v) decision support. Each phase contains a chain of processes which are executed by a series of supporting application tools. Data sourcing consists of two stages: (i) data integration and (ii) data warehouse. Data analysis includes five crucial tools: Online analytical processing (OLAP), ad-hoc analysis, data mining, predictive analysis, and customised and in-house tools. Situation awareness (SA) helps decision makers understand the impact of influential spatio-temporal business factors, including environmental components and events, and thereby, assists them in making quicker and better decisions in the maze of big data sources. Three components of SA are (i) states (actual awareness of environmental components and events), (ii) systems (distribution of SA and their interactions) and (iii) processes (dynamic changes of SA states and their main causes). After assessing the impact of situational factors, BI analyses the hazardous factors and prioritizes them. Next, BI determines the risk exposure of all relevant factors by considering the likelihood of risk and the economic impact associated with each factor. Based on the risk exposures, business objectives, opportunities and affordability, resource availability and other relevant issues and information, a decision support system produces decisions including alternatives that can improve operational efficiencies and services.

To improve the effectiveness of a business system, as mentioned in Section 2.1.2, a number of research studies have shown that the application of contextual information to a business system can bring broad and clear views about the business situation. Therefore, it can help an enterprise to make better decisions [109–111]. The contextual information is mainly considered in the first stage (data integration) of data sourcing and in some of the tools includes ad-hoc analysis for the data analysis phase of BI systems. However, context awareness is not strongly focused in the data warehouse, although a

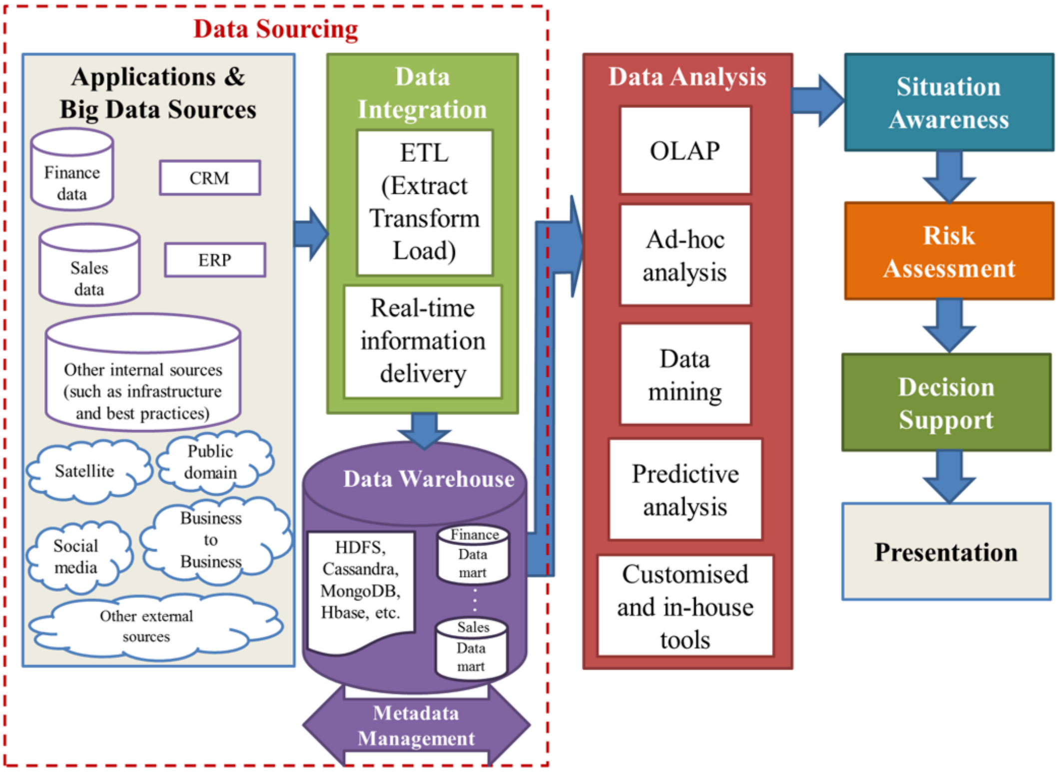

Figure 2.8: Typical business intelligence framework in big data era

few approaches exist to metadata management that give access to information on the business context.

The most challenging mission of context-aware BI systems is how to extract useful contextual information from a variety sources and integrate it. Differences in the nature of data from diverse sources lead to difficulty not only in integrating them but also making them usable and valuable in BI systems. Current BI research studies claim that a large number of contextual information has been considered in their methods. However, this information is mainly collected from internal sources, not from external sources such as social media and the public domain [87]. For instance, the information on user context is gathered from customer data of customer relationship management via information retrieval.

As stated previously, the application of BI is not only limited to business sectors but also to any other organisation. This leads to a broad range of contextual information which influences the key performance indicators [111,112]. For example, BI systems in industry normally concentrate on the customer (user) context (e.g., user background, user interests), the competitor (market) context (e.g., new trends, new products) and the environment (surrounding business) context (e.g., supplier, economic event). In contrast, health care sectors are interested in the environmental context (e.g., weather, temperature, disease), and government organisations give more attention to the national economy context (e.g., population, education, investment, GDP) and the global context (e.g., global economic crisis, the effect of the US presidential election, Brexit in 2016). User, environment and query contexts are the most popular contexts for any field. However, the advent and importance of big data and the Internet of Things have attracted researchers to consider two more contexts for BI: (i) geospatial context, and (ii) business context.

(i) Geospatial context The geospatial context mostly refers to an objects location such as store, customer and partner address locations. This information can be easily found from internal and external databases (e.g., the customers address in the customer database, the suppliers address by exchanging information in the B2B model). Geospatial context is one of the most important information that BI systems need to consider. For example, instead of seeing business figures in separate tables, a manager can have a broad and comparative view of different locations in a single image.

(ii) Business context Unlike the geospatial context, the business context normally refers to the meaning of the content of business data, policies that govern the business data, and any object or action that participates in or reflects business activities [87]. The concept of involving the contextual information of a business contained in databases was introduced in the 1990s, such as adding business rules to databases [113]. In addition to business rules, objects (e.g., staff, customer, supplier, product), and activities (e.g., buying, refunding, giving feedback) are also considered to be the business context [87, 114].

There exist a few context-aware methods on the first two phases of BI, namely data sourcing and data analysis, which we describe below.

2.1.5.1 Data sourcing

Data sourcing comprises two stages (i) data integration and (ii) data warehouse.

(i) Data integration

The quality of data can be high or low, depending on the data collection and aggregation method. As a result, this is the most important aspect of BI systems as it decides whether a BI system works effectively or not. In this

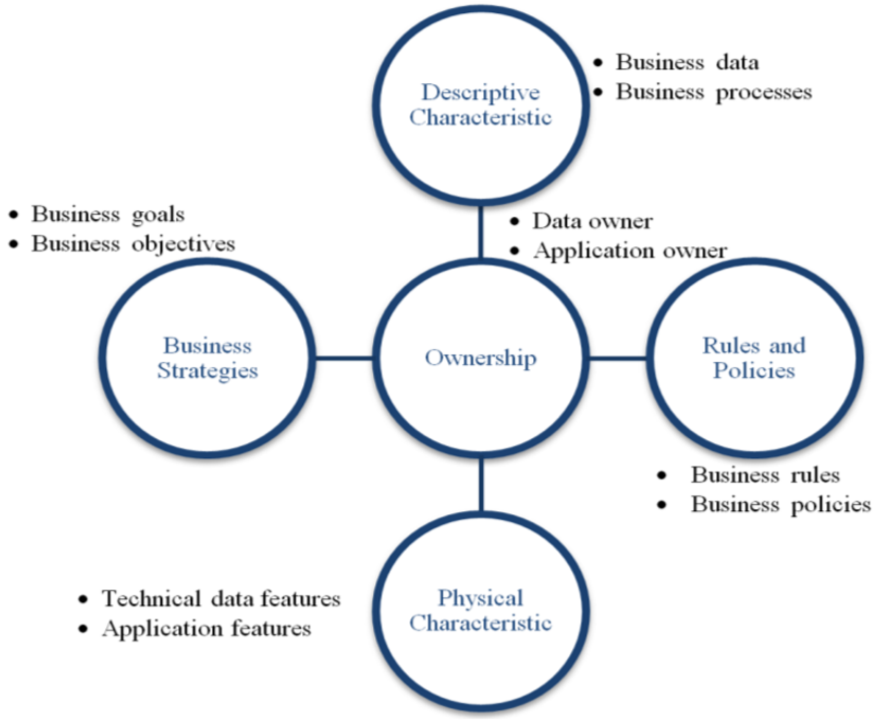

Figure 2.9: Classification in business metadata

stage, not only extract-transform-load (ETL) but also real-time information delivery is critical. The profound need for big data and real-time processing has brought the concept of "real-time information delivery". Real-time information delivery is detailed in the next section.

Extraction of information from various data sources and applications relies on the tools that support the ETL stage. For example, adopting user context into information retrieval has been proven to yield better results for data extraction [12, 70]. In addition to [65], another study by Berlanga & Nebot [87] shows that it can also capture relevant context data from not only internal but also external sources. The approach in this study also helps BI systems deal with integrating internal and external data with each other [87].

(ii) Data warehouse In this stage, context awareness is focused in metadata management. The concept of metadata management has facilitated the way of storing data. However, to our knowledge, to date there have been very few studies in this field. Most recent methods came from information technology companies, and these methods were developed based on their experiences and the situation of customer systems using methods such as interviews and surveys. The first research study on metadata management was published in 1996 [114]. Metadata provide the business context in which business data are used, and can be viewed as a semantic (interpretive) layer of the BI decision-support environment. This semantic layer helps business people navigate through the BI target databases where the business data reside. Metadata also help the technicians manage the BI target databases as well as the BI applications [115].

In [115], metadata are defined as business terms (e.g., the description of common terms in business, such as interest, profit, gross margin), business activities (e.g., buying a product, obtaining customer feedback, marketing a new product) and business policies (e.g., working policies, promotion policies). These activities, objects, strategies and relationships/rules provide the context in which business people use business data. Metadata classifications based on business contextual information are shown in Figure 2.9.

2.1.5.2 Data analysis

While data sourcing mainly focuses on capturing, integrating and storing data, the data analysis phase helps data to become valuable and meaningful by analysing them.

First, data from the warehouse are quickly analysed based on multidimensional cubes via online analytic processing (OLAP) [116]. For instance, a report relating to sales in a chain of retail shops requires the inclusion of five information dimensions: product name, sales volume, store location, date, and store manager. These information dimensions are extracted from the data warehouse by OLAP. Secondly, ad-hoc analysis provides a fast answer to a simple business question. This is done according to the results coming from OLAP by drilling deeper into a business report and then obtaining more details about an object or statistic (e.g., in the previous example, a store manager can have more information about the sales status of a specific store by clicking on the store location).

Apart from the general purposes of this phase, such as organising information for reporting and presenting results based on multiple dimensions, data mining is regarded as the heart of BI because it extracts useful intelligence from data. The tools in data mining that are developed based on several concepts (e.g., classification, clustering, association rule and prediction) are helpful for obtaining deep insight into data by discovering particular patterns and correlations among the data. For example, a data mining tool can give a future sales prediction by analysing similar patterns of sales within a period of time.

Contextual information is normally considered in OLAP and data mining. As stated earlier, the geospatial context helps enterprises to acquire a broad view about sales status across location and area [117]. This context type has been used since the mid-2000s [118] via OLAP cubes. A geospatial context is a result of combining geographic information systems and image processing. Moreover, geospatial data has become well-known in mass-market applications, such as web mapping and location-based services [119].

To support big data mining, high-performance computing platforms (e.g., Hadoop, Mahout) and advanced data mining methods are required [120–122]. Advancements required in data mining for BI include efficiency improvement of single-source knowledge discovery methods, designing a data mining mechanism from a multisource perspective, the study of dynamic data mining methods and the analysis of stream data. Extensive research has been conducted on a variety of topics in data mining, such as mining opinions on social media [123, 124] and optimizing product configurations using data mining approaches [125, 126]. On the other hand, due to the importance and potential benefits of big data mining, business organisations need to adopt filtering mechanisms to enable them to collect efficiently those data that are most meaningful from a business perspective and analyse them according to their requirements. This demands consideration of the business context in big data gathering and analysis. Furthermore, this represents a significant research challenge to the data science community to manage and provide smart and instant data analytics for business decision-making [87].

The following section describes the tools and platforms that support context-based methods and real- time data analytics.

2.1.6 Context awareness in big data tools and platforms for real-time analytics

Two important questions in this regard are: (i) how can big data tools (e.g., Splice machine, MarkLogic, Google charts, SAP inMemory, Cambridge semantics, MongoDB, Pentaho, Talend, Tableau, Splunk) and platforms (e.g., Lumify, Talend Open Studio for Big Data, HPCC Systems Big Data, Apache Storm, Apache Drill, Apache Samoa, Ikanow) implement context-based methods? and (ii) how do these tools and platforms support context-aware methods for real-time big data capture and analytics? For the former, in the literature there exist a number of context-based methods that have been implemented using big data tools and platforms. For example, context awareness has been applied to Hadoop [45], IBM Watson Analytics [127], Microsoft Cortana [128] and SparkContext [129].

The term MapReduce has been in existence for a long time, since even before big data became one of the major concerns in the research community. Since its introduction in the late 1990s, Hadoop has become the major leverage for the MapReduce algorithm to do parallel and distributed processing. Hadoop does this by splitting up the tasks across multiple processing nodes and then recombining them to extract the final results. Much research has been done in this area to find ways to improve the ability to collect and analyse big data by using Hadoop and MapReduce in combination, context-awareness being one. The project PER-MARE (Pervasive Map-Reduce) emphasises that one of its main goals is context-awareness and consideration of context has become an important key in deploying applications successfully in pervasive environments [60, 130]. Although cloud infrastructures have become very popular, pervasive grids are still important for several enterprises which hesitate to use the cloud because of the sensitivity their data and prefer to use their internal resources. The PER-MARE project aims to adapt Hadoop to pervasive environments and add context-aware behaviour via a lightweight context middleware [130].

With the purpose of embedding context-aware support for the MapReduce and Hadoop applications, a set of important requirements are divided into three main sections: (i) general requirements, (ii) context modelling requirements and (iii) distribution requirements. The contextual information required includes the status of pervasive grid resources and new processing nodes that can be integrated in the pervasive grid at any moment. The information regarding these contexts mainly supports scheduling tasks in order to distribute the tasks among available resources, scheduling being an important service for grid platforms. Due to heterogeneous resources and the dynamically changing environment, it is important for better scheduling to consider current contextual information when deciding task distributions [60]. Cassales et al. [131] also introduced a scheduling scheme for Hadoop, based on the concept of context awareness in a pervasive environment. While the PER-MARE project developed a lightweight context middleware to provide the Hadoop platform and its applications with context information, Cassales et al. [131] aimed to embed context-aware ability into the Hadoop task-scheduling scheme through a context collector on the Hadoops resource manager. Contextual information is collected by schedulers to detect dynamic changes with the aim of finding the available resources on the nodes. In addition, slave nodes must communicate periodically with the master to keep information updated and allow the scheduler to adapt to the new context [132]. The PER-MARE project and the research by Cassales et al. were developed to schedule the tasks among resources, particularly for pervasive grids. However, Hadoop has also been developed as a big data cloud framework and it can handle huge amounts of event data. However, as Hadoop is a batch-oriented data processing framework, this leads to high latency, such as the inability to respond immediately in emergency situations. To solve this issue, an existing ontology-enabled collaboration framework, which uses contextual information, is combined with a Hadoop cloud framework to build a complementary cooperating system [133]. On the other hand, a number of studies in this area focus on building smart cities based on cloud-based context-aware information services, such as an architecture-based cloud for context-aware citizen services in [134].

Furthermore, the advancement of technology, including mobile devices and sensors as data sources, the instant availability of data through faster mobile broadband and cloud services, the quest of businesses to have real-time monitoring systems and the expectation of having business intelligence instantly at their fingertips using ad-hoc queries have accelerated the demand for real-time big data analytics. A number of platforms and tools that have been introduced for real-time data analytics are listed in Table 2.11 with their supported file systems and latency and throughput [135, 136].

As a consequence, business organisations have pushed the relevant research community and vendors to introduce context-aware real-time big data analytic tools and platforms. For example, refer to the tools and platforms introduced for context-aware stream computing in IBM InfoSphere Streams [137]. Building contexts and profiles of entities and connecting streaming data with contextual information are possible. Several additions to the Geospatial Toolkit and the Time Series Toolkit in IBM InfoSphere Streams (version 4.0) also help provide context-aware stream processing, enabling greater insights in making superior real-time decisions [137]. In addition to these context-aware real-time business analytics, context-aware real-time analytics are also considered in video capture and dissemination. To improve the ability of searching in real-time through the video repositories and to increase their worth, context information extracted from sensor data (e.g., time, location, speed, temperature, heartrate) and other sources such as information extracted through video analytics are added to the videos label. For instance, the study reported in [138] has shown that capturing contextual information as above along with the capacity to search any video subset selected based on deep context-based concepts, can create a valuable video public resource.

Although there are a number of techniques in the current literature, measuring the accuracy and trustworthiness of the contextual information provided by a context provider remains a major research challenge to the big data research community which is yet to be resolved.

Table 2.11: Popular tools and platforms for supporting real-time analytics

Name	Description	File system	Latency	Throughput
Hadoop	A framework for distributed processing of large data sets across clusters of computers using simple programming models.	HDFS (Hadoop Distributed File System)	Depends on the tools (e.g., Spark, Kafka) developed in Hadoop environment; in general it is high	Depends on the tools developed in Hadoop, but in general throughput is high
Apache Flink	Streaming dataflow engine	HDFS, Amazon S3, MapR file system, Tachyon	Low (configurable)	High
Apache Spark	Large-scale data and batch processing engine (enables emulated streaming via micro-batching)	HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat as below: Version 1.5.x and older: distributed dataset, version 1.6.x and later: distributed dataset and (key, value)	Medium (depending on batch size)	High
Apache Storm	A system for processing streaming data in real time reliably through enterprise Hadoop platform	MongoDB, RDBMs, Cassandra.	Very low	Low

Continued on next page

Table 2.11 – continued from previous page

Name	Description	File system	Latency	Throughput
Apache Samza	A distributed stream data processing framework	Apache Kafka for messaging. Apache Hadoop YARN to provide fault tolerance, processor isolation, security, and resource management.	Low	High
Apache Apex	Enterprise-grade unified stream and batch processing engine	HDFS, S3, NFS, FTP, Oracle, MySQL, Cassandra, mongoDB	Low	High
Apache Kafka	Distributed, fault tolerant, high throughput pub/sub messaging system	HBase, Cassandra, Storm, RDBMS	Very low	High
Flume	A distributed streaming system that supports reliability and availability for big data.	HDFS, HBase, Cassandra	Low (for streaming)	High (for streaming)
SAS Event Stream Processing (ESP)	Stream data processing from operations, transactions, sensors and devices in real time.	HDFS, IBM DB2, SAP Sybase ASE, SAS data sets and others.	Low	High
SAP HANA	Real time analytics	SAP HANA database	Low	High

Continued on next page

Table 2.11 – continued from previous page

Name	Description	File system	Latency	Throughput
Oracle Event Processing	A complete solution for building applications to filter, correlate and process events in real time	HDFS, NoSQL, Java Message Service (JMS) systems, and caches.	Medium	High
Microsoft Azure	Stream analytics, real-time stream processing in the cloud	SMB 3.0 protocol	Low	High
InfoSphere Streams	An advanced analytic platform	HDFS, Cassandra, MongoDB, Spark	Low	High

2.2 Business Process and Business Strategy

According to traditional thinking, effective strategic plans help enterprises to achieve their goals. However, in an era of intense competition, great strategic plans are not sufficient for enterprises to achieve their goals. While business goals are long-term, business strategies are more specific and they help enterprises to have a better view of what to do. To achieve the goals in the strategic plans, enterprises need to know what to do and where (or when) to start based on their strategies. Therefore, transferring strategies to specific actions which are well defined in business processes is one of the most important steps for strategy deployment.

In general, business strategies can be divided into three main groups (i) strategies for the improvement of an existing system (e.g., a strategy for increasing the number of customers using shopping online services), (ii) strategies for repairing loopholes in the existing system (e.g., a strategy for fixing security loopholes in banking system) and (iii) the strategies for a new investment (e.g., a strategy for a new product launch). As these examples show, each strategy includes the expected outcomes. These outcomes are achieved by executing a list of actions and tasks which come from business processes. A business process typically includes a series of coordinated events (or activities) across a number of functional elements of an enterprise and it is executed by stakeholders (or customers).

With the purpose of improving the connectivity between strategies and processes, the term "strategic alignment" has been defined, and it has attracted the attention of a large number of researchers from the relevant research community for many years. Strategic alignment may be defined broadly and it is relevant to many aspects of an enterprise [139]. However, it is mainly defined based on two points of view. Firstly, strategic alignment refers to how information technology (IT) aligns with business needs [140] (e.g., how to develop IT systems and determine their roles in order to suit business strategies). Secondly, strategic alignment is also seen as a technique where the bridge communication between strategies and processes is built and described [141]. Although, these two definitions sound different, they complement each other and there are several important common aspects. That is, business strategies normally come from business departments, whereas business processes are generally run with the support of IT. The strategic alignment mentioned in this chapter mainly concentrates on the second view, i.e., the coherence between strategies and process models. To avoid confusion, in this thesis, the term "strategy-process relationship" is used instead "strategic alignment".

Business strategies and processes are the core elements in any enterprise in order to achieve their goals. The goals will not be able to be reached if the business processes do not reflect business strategies, and vice versa. A number of studies have shown that a strong connection between strategies and processes can bring huge advantages to an enterprise [142–144]. For example, to achieve the goals of increased customer satisfaction and reduced number of staff, a part of the business processes for customer service from Jetstar Airways (jetstar.com) has been transferred from manual jobs to automated processes such as virtual staff working on their website. These automated processes allow Jetstars customers to find suitable answers for their queries via virtual staff without waiting in a long queue to talk to real staff. As a result, instead of letting their customers wait for a long time to receive a response, these automated processes help Jetstar Airways not only reduce the number of staff working on telephones, but also to increase customer satisfaction as their customers are able to find the information they need within an acceptable amount of time. In an other example, in order to increase the accuracy of aircraft control and reduce the stress of pilots during flights, especially in long-haul services, Boeing and Airbus have embedded process automation into their planes with the purpose of supporting pilots during take-off and landing [145].

To ensure that all processes run smoothly whilst still satisfying a number of business strategies, the main objective is that business processes need to be appropriate and up-to-date so that they fit accordingly. As mentioned above, a business process typically contains a series of (automatic and manual) activities and interactions between stakeholders and customers. This leads to complexity in managing them. Therefore, executing them under the roof of business strategies and finding the loopholes in the strategies of regulatory compliance are always difficult.

Although a number of studies and techniques have been introduced in order to help enterprises transferring from strategies to processes, gaps remain in the interaction between business strategies and processes. The results and findings that are claimed from the studies and those collected from practices normally do not match [146,147]. In addition, they are not sufficiently flexible for application in different systems. This is because each methodology was developed for a specific individual system. Furthermore, most still rely on the concept of string comparison. Several studies have claimed that their methodologies have improved the correctness of process models by analysing business rules [141, 148]. However, the rules used in their implementation are limited. As a consequence, no evidence has been provided that the methodologies still produced the expected outcomes in complicated practical business environments. The flexibility and adaptability of these studies are low, because they are only suitable for application to selected scenarios. As a consequence, this leads to the inability to adapt to the rapid changes and the complexity of real business environments. Uncertainty and imprecision in defining the effect of a process may make the system unsuitable for real world applications. To reduce this uncertainty, establishing the relationship between processes and strategies based on rule-based inference models appears to be one of the most suitable techniques for business.

To improve the effectiveness of the connection between business strategies and processes, all studies in this field have focused on the three main objectives:

(i) To clarify which business process needs to be involved once a business strategy starts running.

(ii) To avoid developing duplicated business processes.

(iii) To examine if a business process is following regulatory compliance and strategy.

When a new business strategy is released, a list of business processes is assigned in order to execute the tasks that the success of the business strategy. Which business processes should be included in this list? Is it necessary to develop a new process if an approximately similar process is found in the enterprise system? To answer these questions, objectives (i) and (ii) have attracted the research community the most.

To choose the most appropriate business processes and to avoid duplicated business process development, a large number of studies have been conducted. Semantic techniques have arisen as the most popular solutions to achieve all three objectives. According to the concept of semantic techniques, each process is analysed and tagged with several semantic labels. These labels may refer to the tasks for which the process is responsible and the expected outcomes of the process. These labels are developed depending on the graphical languages of semantic techniques and they play a role in communication between business strategies and processes [149, 150]. Semantic techniques for process labels also help to handle compliance gap detection. However, applying semantic techniques for process labels suffers from several drawbacks. For example, finding the relevance between business strategies and processes, and detecting the gaps in regulatory compliance need to go through the semantic labels, the execution can take time without giving a correct outcome if the semantic labels belong to businesses processes or a regulatory compliance are not defined [151].

As semantic techniques for label models have shortcomings, researchers have tried another way to solve this issue. Therefore, the concept of adding annotation into processes has been proposed and its advantages have been shown by studies and practice [151, 152]. For example, Governatori et al. [153] found that objective (iii) can be achieve by process formalization based on semantic annotations. Fellman et al. [151, 154] also stated that each annotation in the process model is associated with a selection of formal ontologies. In this case, the regulations are represented by formal ontologies. The main purpose of developing the relationship between them is not only to check the correctness of the model, but also to ensure that the model evolution is appropriate [151, 154].

However, considering only those elements which belong to process models may lead to lack of capture of business rules [148]. Therefore, several studies have focused on embedding business rules in the relationship between business processes and strategies [141, 155]. In [141], business rules are executed based on the annotations of processes and a set of strategies. These rules are then used as a buffer step with the aim of minimizing the set of business processes, in order to achieve a particular strategy. Alternatively, regulations are transformed and computerized into rules. Annotated process models are then checked via the connection between the annotation of process and the representation of rules [155]. The difficulties in managing the collections of processes increase significantly due to the lack of associated tools. Therefore, [118] introduced a multi-layered approach in order to make a link between strategies and process models.

The examples described above show that it is important to find a relationship between business strategies and processes. A business strategy can only be successful if a company knows where to start their plan, where to improve and invest. Therefore, a rule-based inference model is essential, as it allows the incorporation of rules by domain experts and it can handle uncertainty in an effective and efficient manner. To our knowledge, there exists no such rule-based inference model to find which processes relate to which strategies in this area.

2.3 Semantic Similarity for Big Data Analytics

Because of the complex structure and the variety of data sources of big data, the application of semantic manipulation during big data processing has shown improvement in the results of big data analytics [156]. A number of approaches to semantic similarity calculation have been proposed in the literature. However, in this section, we focus on a review of the approach that has gained popularity and citations in the research area.

According to notable semantic knowledge bases such as WordNet (a General English Language Usage) [157] and Gene Ontology [158], they are developed depending on a hierarchical structure and built for specific human purposes. Therefore, for linguistic semantic aspect, the semantic value for each semantic word is automatically calculated based on this concept.

Determination of the semantic similarity between words is helped by the analysis of the hierarchical structure of the language knowledge base. In WordNet, each word belongs to a synonym set called a synset, and the shortest path between two words means the minimum distance connecting the synsets of these two words. For example, the shortest path between "cabbage" and "orange" is "cabbage" - "vegetable" - "product" - "fruit" - "orange" as shown in Figure 2.10. Therefore, the minimum distance between "cabbage" and "orange" is 4. In addition, the synset of "product" is called the subsumer for "cabbage" and "orange".

In the hierarchical structure of the language knowledge base, given a keyword called k and its semantic word w, the semantic similarity value abbreviated $\xi(k,w)$ between these two words can be calculated based on the shortest path between them, l. If k and w are in the same synonym set (synset) then the shortest path between them is 0. If they are not in the same synset but there are one or more similar words in their synsets, then the shortest path is 1. In all cases, the semantic similarity value between two words is calculated as follows:

$$\xi(k,w) = e^{-\alpha l}$$

with $\alpha = 0.2$ [159]

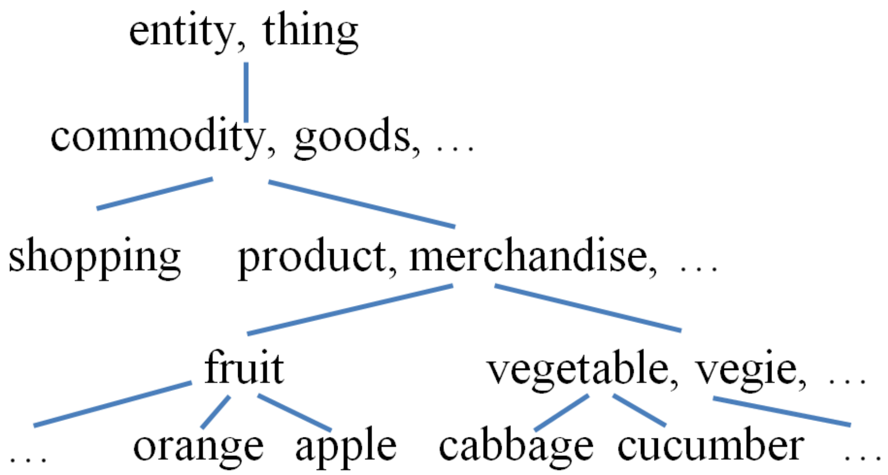

Figure 2.10: An example of a semantic hierarchy

However, only considering minimum distance between words may generate a less accurate result [159]. For instance,assume there are three words A, B and C. If the distance between A and C is shorter than the distance from A to B in the language knowledge base, then the word A is more similar to the word C. However, this frequently happens in a huge knowledge base. Nevertheless, this leads to the fact that the meaning of the words A and C sometimes is not as close as the meaning of A and B.

Therefore, more features need to be considered and added to improve the accuracy of the calculation of semantic similarity between words, in this case, the hierarchy distance. A hierarchy distance indicates the depth or distance from the root of the knowledge base to the considered words in the ontology in WordNet. Normally, a hierarchical structure consists of many layers and the words presented at upper layers are more general and less detailed than the words at lower layers. As a consequence, the semantic similarity between the words at the upper layers are also less than those in the lower layers.

When the hierarchy distance h between k and w is added to the semantic similarity calculation, the formula of $\xi(k,w)$ is changed as shown below:

$$\xi(k,w) = e^{-\alpha l} \times \frac{(e^{\beta h} - e^{-\beta h})}{(e^{\beta h} + e^{-\beta h})} \tag{2.1}$$

Where,
l = shortest distance between k and w
h = hierarchy distance between k and w
α = 0.2 and β = 0.45 [159]

2.4 Research Challenges

As emphasised in this chapter, the importance of semantic manipulation to capture deep insights (see Section 2.3) and embedding context into computing technologies to support business management systems, and the effectiveness of these methods/applications has been recognized by the relevant research community. Although many approaches/applications using semantic information and context models have been developed, major challenges for business organisations and significant research issues for the research community remain. Future research challenges in semantic manipulation and context-aware systems for business applications are outlined below:

1. If an organisation needs to capture deep insights and real-time or quasi-real-time processing at the same time, we need to make a trade-off between semantic manipulation and computational processing time for both data collection and data analytics. This requires the scaling of the amount of semantic manipulation considering business requirements.

2. Current context models lack the use of image, audio and video data to represent themselves. However, in addition to text data, these multimedia elements can play a vital role in depicting a clearer picture of a context. To date,

the techniques for metadata representation defining contexts in the domain of these elements and the scalability of those techniques to capture instantaneous context adaptation have not been explored. Therefore, challenges lie in developing innovative context-aware methods to extract comprehensive contextual information from these diverse data types.

In addition, many metadata repository products are still designed by technical experts rather than business experts. Some of these products still have a cryptic metadata language, lack sophisticated reporting capabilities, are not context-sensitive, and require an understanding of the meta-model that describes the metadata objects and their relationships. This indicates the pressing need for the development of a meta-data model focusing on a business view and context, in order to reduce the requirement to understand metadata objects and their relationships.

On the other hand, the role of a metadata repository is extremely important in BI systems. A BI needs better metadata repository support to manage databases, as well as methodologies or techniques that can analyse and process data more effectively. For example, as mentioned previously, unlike IR, which collects only raw data, the main target of BI is to capture deep insights from the data or the value of the data. Beyond the hype of big data, business intelligence needs to overcome these challenges and difficulties.

3. A business organisation needs real-time processing of big data so that a user receives the required and desired information on time. The studies presented in Section 2.1.6 highlight the importance of contextual information in real-time big data analytics, but the effort is still in an embryonic stage. The difficulty lies in the filtering and validation of relevant contextual information and processing within the time limit specified by the respective application.

This motivates the rigorous investigation of real-time context-aware big data processing techniques and the instant manipulation of contextual information of any business entity, including internal and external entities.

2.5 Conclusions

In this chapter, an overview of context-based methods in big data analytics for business applications is presented. A general background to big data analytics, the definition of context, context models and their evolution, and context evaluation techniques are also provided. Context-aware big data analytic techniques are divided into four main groups: (i) context-based methods in business process and workflow, (ii) context-aware information retrieval, recommendation systems and query classifications, (iii) context-based methods for business intelligence, (iv) big data tools and platforms for context-aware and real-time analytics. In each group, the contextual information used and the purpose of considering context are highlighted. Furthermore, the role of business strategy, business process and the relationship between them are discussed. In addition, the approach to semantic similarity calculation is also described.

The state-of-the-art methods for each context-aware big data analytic group and the methods of semantic manipulation and relation between business process and strategy are reviewed, illustrating their main purposes, advantages and drawbacks. A number of further research challenges and their significance in business applications are emphasised.

Based on these research challenges, in this thesis, a number of research objectives have been identified, as presented in Section 1.3 of Chapter 1. The next chapter introduces a new approach, namely "A Rule-Based Inference Model for Establishing Process-Strategy Relationship" which will present a significant component of the research project required to address research Objective 1.

Chapter 3

A Rule-Based Inference Model for Establishing Process-Strategy Relationships

To reduce uncertainty and accurately determine the relationship between business processes and their relevant strategies as defined by business domain experts, and hence to present the research approach required to fulfil Objective 1 presented in Section 1.3 of Chapter 1, this chapter introduces a proposed rule-based inference model. While "rule-based" heuristics have been applied to a number of applications, they have not been considered in regard to discovering the relationship between business processes and strategies. Our model will not only help business organisations to achieve their goals by discovering which business processes need to be involved in which business strategies, but will also reduce the possibility of losing important details of business process optimisation.

The proposed inference model introduced in this chapter is developed with the purpose of finding a business strategy and its relevant processes based on a rule-based inferencing system. This is developed from the relevant facts and annotations. A set of business strategies and processes are related to each other only if they satisfy the rules defined in the knowledge base. We have developed a business case to validate our proposed model and the results show that our model can infer the relationship accurately using the knowledge base defined for the business case.

The structure of this chapter is as follows: the rule-based inference model is introduced in Section 3.1. The case study for the proposed inference model is described in Section 3.2 and finally the conclusion is presented in Section 3.3.

3.1 Proposed Rule-based Inference Model

Because many loopholes still exist in the current studies, as stated above and in Chapter 2, the methodology presented here is a breakthrough in business process management. The proposed inference model is the first of its kind in managing processes and strategies based on rules in the knowledge base. The model will enable a business to determine the fulfilment of objectives considering the activities executed through its business processes.

The purposes of the model are:

- to find a business strategy and its relevant processes based on rule-based inference developed from facts and annotations.
- to generate the list of strategies and processes based their facts and annotations, respectively that make given rules to satisfy.

Figure 3.1 shows a schematic diagram of the inference model. As shown in the figure, the inference model obtains the strategies with their corresponding facts (see Table 3.1) and the processes with their corresponding annotations (see Table 3.2). Annotations are the information on a business process, and facts are the results of a strategy. A process may belong to many strategies

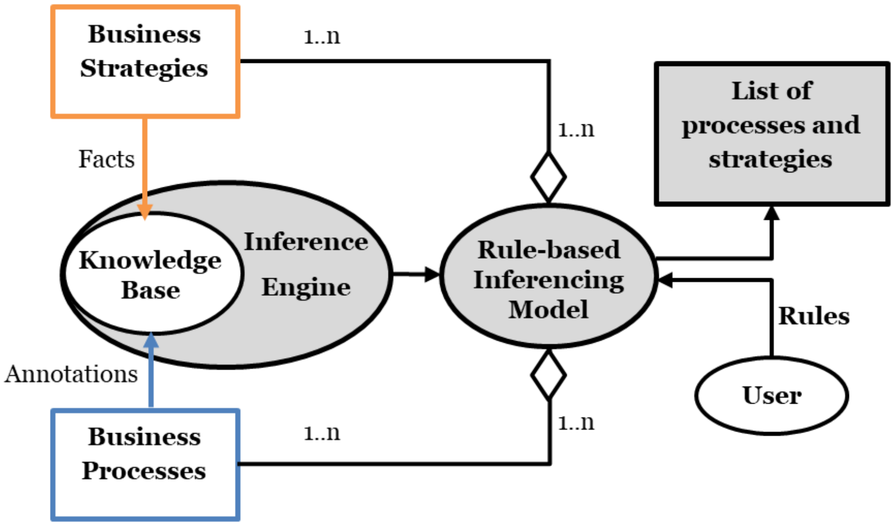

Figure 3.1: Schematic diagram of proposed inference model

and a strategy may need many processes to be executed. Using the reasoning process, the proposed inference model produces the processes and their relevant strategies. As expected, the inference results establish many-to-many relationships between processes and its strategies.

Table 3.1: Processes and their annotations

Process	Annotation
Pricing (P1)	FinalPrice
Business planning (P2)	ProductPlanning SupplierPolicy StorePlanning StaffPolicyDeveloping SafetyPolicyExecute InternshipProgramDeveloping
Inventory management (P3)	ProductPlanning InventoryPlanning

Continued on next page

Table 3.1 – continued from previous page

Process	**Annotation**
	StorePlanning
Product scheduling/planning (P4)	ProductPlanning
Product design/development (P5)	ProductPlanning
Quality management (P6)	StorePlanning ProductPlanning
Customer service (P7)	Feedback Membership Communication Refund
Invoicing/Billing (P8)	Invoice
Sales/Order entry (P9)	OrderStatus Delivery
Advertising/Promotion (P10)	ProductPlanning PromotionPlanning
Distribution (P11)	ProductPlanning StorePlanning
Purchasing (P12)	SupplierPolicy
Marketing research (P13)	ProductPlanning StorePlanning
Personnel management (P14)	StaffPolicyDeveloping SafetyPolicyExecute InternshipProgramDeveloping *SalaryPayment* *RecruitmentPlanning*

Table 3.2: Strategies and their facts

Strategy	Fact
Lowering the cost of planning (S1)	HappyCustomer LowRetailPrice LowProductionCost
Maintaining quality of own brands (S2)	HappyCustomer LowRetailPrice GoodService SufficientQuantity ProductQuality Promotion
Building strong, collaborative partnership with suppliers (S3)	ReliableSupplier
Long-term partnership (S4)	ReliableSupplier
Offering fresh product to customers (S5)	HappyCustomer ProductQuality
Improve the quality and availability of fresh food (S6)	HappyCustomer ProductQuality SufficientQuantity
Long standing suppliers (S7)	ProductQuality SufficientQuantity LowWholesalePrice LowProductionCost ReliableSupplier
Develop an innovative and unique product (S8)	HappyCustomer ProductQuality LowRetailPrice Promotion

Continued on next page

Table 3.2 – continued from previous page

Strategy	Fact
Improving store network across all own brands (S9)	HappyCustomer OrderPolicy ConvinientStore StoreDecoration
Making shopping easier and simpler for customers (S10)	HappyCustomer OrderPolicy ConvinientStore StoreDecoration
Creating bigger, brighter stores with new features (S11)	HappyCustomer OrderPolicy ConvinientStore StoreDecoration
New look for Liquorland stores (S12)	HappyCustomer OrderPolicy ConvinientStore StoreDecoration
Opening first online standalone store (S13)	HappyCustomer NewChannelsServices
Expanding into new channels and services (S14)	HappyCustomer NewChannelsServices
Bringing together diverse backgrounds (S15)	HappyStaff StaffPolicy SafetyPolicy InternshipProgram
Launching first accessibility action plan (S16)	HappyStaff StaffPolicy

Continued on next page

Table 3.2 – continued from previous page

Strategy	Fact
	SafetyPolicy InternshipProgram
Increasing the number of women in leadership (S17)	HappyStaff StaffPolicy SafetyPolicy InternshipProgram
Supporting indigenous team members (S18)	HappyStaff StaffPolicy SafetyPolicy InternshipProgram
Maintaining a safe workplace (S19)	HappyStaff StaffPolicy SafetyPolicy InternshipProgram
Maintaining graduate program (S20)	HappyStaff StaffPolicy SafetyPolicy InternshipProgram
Natural refrigerants (S21)	ProductQuality SavingEnergy
Growing exclusive brands (S22)	HappyCustomer ProductQuality LowRetailPrice
More bonuses instead of salary increase (S23)	*HappyStaff* *StaffPolicy* *IncentiveBonusPolicy*

A key action or an effect of a process is represented by an annotation, whereas, facts express strategies in brief terms. Note that, a process may contain many annotations and a strategy may have many facts. Each annotation may belong to more than one process, and each fact may belong to more than one strategy. When a stakeholder gives an input which has two parts (i) antecedent condition (including annotations connected by logical operators) and (ii) consequent condition (including facts connected by logical operators), a list of strategies and their relevant processes is returned, based on rule-based inference execution. All these steps are required to work on a collection of processes and strategies which are provided from the knowledge base, as shown in Figure 3.1 and Algorithm 3.1.

Algorithm 3.1 Inference process-strategy relationship

Input: $R : ac \implies fc$

ac = Antecedent condition of rule R involving annotations and logical operators
fc = Consequent condition of rule R involving facts and logical operators
a = List of annotations = $[a_1, a_2, a_3, \cdots, a_n]$
f = List of facts = $[f_1, f_2, f_3, \cdots, f_m]$
p = List of business processes = $[p_1, p_2, p_3, \cdots, p_k]$
s = List of business strategies = $[s_1, s_2, s_3, \cdots, s_l]$

if $(\neg ac \vee fc)$
    use process_of (p_i, a_i)
    when
        for each a_i in the given annotation list a
            annotation_of (a_i, p_i)
    use strategy_of (s_j, f_j)
    when
        for each f_j in the given annotation list f
            fact_of (f_j, s_j)
    use inference_process_strategy (p_i, s_j)
    when
        for each p_i of process_of (p_i, a_i)
        for each s_j of strategy_of (s_j, f_j)

In Algorithm 3.1, ac is the antecedent condition containing a list of annotations and fc is the consequence condition including a list of facts. The elements in each condition are combined with logical operators, such as $\wedge$ and $\vee$ which represent AND and OR respectively. Based on theoretical principles, the rule $R : ac \implies fc$ is not satisfied only if the antecedent condition outputs TRUE whereas the consequent condition returns FALSE. According to logical implication and based on the outcomes of antecedent and consequent conditions, the rule satisfaction decisions are given in Table 3.3. Our proposed system takes the input of a rule, i.e., antecedent and consequent conditions and then evaluates them. It then evaluates the rule (Step 1 of Algorithm 3.1).

In addition, in this algorithm, the process_of (p_i, a_i) is generated based on Step 2 to Step 5. Similarly, the strategy_of (s_j, f_j) is generated based on Step 6 to Step 9. Finally, the inference model generates an inference_process_strategy (p_i, s_j) using the outcomes of process_of (p_i, a_i) and strategy_of (s_j, f_j) (Steps 10 - 13).

Table 3.3: Rule satisfaction decision

Antecedent condition (ac)	**Consequent condition (fc)**	$ac \implies fc$
1 (TRUE)	1 (TRUE)	1 (TRUE)
1 (TRUE)	0 (FALSE)	0 (FALSE)
0 (FALSE)	1 (TRUE)	1 (TRUE)
0 (FALSE)	0 (FALSE)	1 (TRUE)

In Table 3.3, if the antecedent condition returns FALSE or both antecedent and consequent conditions are FALSE, then the rule is still satisfied. This is also considered in our model to execute any rule a user wants, by providing an option to input a rule into the system in the proposed model. The evaluation process of a rule entered by a user is shown in Figure 3.2. The specific example in this figure represents how a rule is satisfied based on the given annotations and facts. The corresponding annotations and facts of antecedent and consequent conditions are replaced by 1 (TRUE) if they are found in the knowledge base; otherwise they are replaced with 0 (FALSE).

As indicated in Algorithm 3.1, the structure of the rule is an implication of antecedent and consequent conditions. The antecedent condition comprises annotations called A_i, whereas the consequent condition comprises facts called F_j. Note that i and j are finite and range from 1 to n for annotations and from 1 to m for facts. The annotations and facts of each side are connected by logical operators such as AND ($\wedge$) and OR ($\vee$). The antecedent condition in the figure is given as $[(A_1 \wedge A_2) \vee A_3]$. It is assumed that A_1 and A_2 are found in the knowledge base, whereas A_3 is not. Therefore, the condition is equivalent to $[(1 \wedge 1) \vee 0]$. By evaluating the condition, it outputs '1'. Similarly for the consequent condition, which is given as $[(F_1 \vee F_2 \vee F_4) \wedge F_3]$. Assume that F_2 and F_3 exist in the knowledge base, while F_1 and F_4 do not. The condition is then equivalent to $[(0 \vee 1 \vee 0) \wedge 1]$, and therefore after evaluation, it returns '1'. As both antecedent and consequent conditions are evaluated as '1', the rule is satisfied. After this, the outcome including a series of processes and strategies is returned.

3.2 Case Study

The above inference model was implemented using the Pyke package (pyke.sourceforge.net) in the Python environment. Pyke is built in the Python environment and is an artificial intelligence tool like Prolog (swi-prolog.org) which can be used to develop an expert system. This helps to develop a knowledge

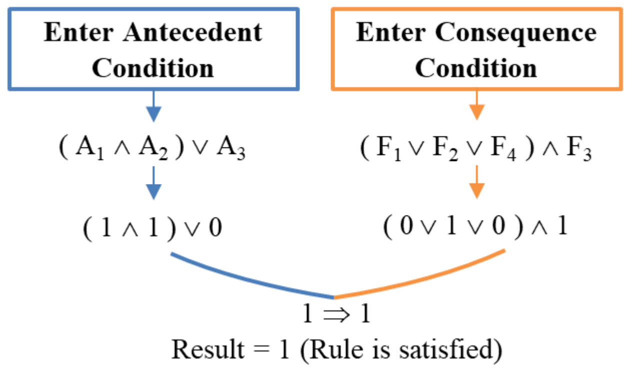

Figure 3.2: Pictorial representation of rule-based inferencing

base and execute the inference process. Our proposed model was validated by considering a business case of a retail supermarket. All the relevant processes of an enterprise computing system and the strategies to meet the business objectives were determined after consultation with the enterprise business experts. As stated in the previous section, a key action or an effect of a process is represented by an annotation, while facts express a strategy in brief terms. The annotations of each process and the facts of each strategy are also developed manually considering the selected enterprise computing and the business organisation, respectively. Table 3.1 shows the processes and their corresponding annotations, while the strategies and their facts are presented in Table 3.2.

For example, the aim of the process 'Pricing' is to set a price for a product or service, so the effect for this process is defined as 'FinalPrice'. Therefore, 'FinalPrice' becomes the annotation of the process 'Pricing' (P1) as shown in Table 3.1. Similarly, the target of the strategy 'Lowering the cost of planning' is to increase customer satisfaction (HappyCustomer), to impose a minimum retail price (LowRetailPrice) and thereby to achieve LowRetailPrice, a cost reduction in manufacturing (LowProductionCost) is needed. Hence 'HappyCustomer', 'LowRetailPrice' and 'LowProductionCost' become facts for strategy (S1) in Table 3.2.

In Table 3.1, a process may contain only one annotation (e.g., the process 'Pricing' has only one annotation 'FinalPrice') or more than one annotation (e.g. the process 'Business planning' has six annotations 'ProductPlanning', 'SupplierPolicy', 'StorePlanning', 'StaffPolicyDeveloping', 'StafetyPolicyExecute', 'InternshipProgramDeveloping'). Similarly, an annotation may belong to only one process (e.g., the annotation 'Invoice' belongs to only one process 'Invoicing/billing') or more than one process (e.g., the annotation 'ProductPlanning' belongs to eight processes 'Business planning', 'Inventory management', 'Product scheduling/planning', 'Product design/development', 'Quality management', 'Advertising/Promotion', 'Distribution', and 'Marketing research').

In a similar way, in Table 3.2, a strategy may comprise only one fact. For example, the strategy 'Long-term partnership' has only the fact 'ReliableSupplier', while the strategy 'Lowering the cost planning' has three facts 'HappyCustomer', 'LowRetailPrice', 'LowProductionCost'. Furthermore, a fact may belong to only one strategy. For example, the fact 'SavingEnergy' belongs to only one strategy 'Natural Refrigerants', while the fact 'LowRetailPrice' belongs to four strategies 'Lowering the cost planning', 'Maintaining great quality of own brands', 'Developing an innovative and unique product', and 'Growing exclusive brands'.

From these annotations and facts given in Tables 3.1 and 3.2, respectively, a knowledge fact base was developed using Pyke, which is shown in Tables 3.4 and 3.5. Table 3.4 shows the knowledge base, namely the annotations_of and their related processes, while Table 3.5 presents the knowledge base, namely facts_of for facts and their corresponding strategies.

Table 3.4: Knowledge fact base (KFB) namely, annotation_of($annotation, $process) for annotations and their relevant processes used in Pyke.

#annotation_of ($annotation, $process)
annotation_of (FinalPrice, P1)
annotation_of (ProductPlanning, P2)
annotation_of (ProductPlanning, P3)
annotation_of (ProductPlanning, P4)
annotation_of (ProductPlanning, P5)
annotation_of (ProductPlanning, P6)
annotation_of (StorePlanning, P6)
annotation_of (Feedback, P7)
annotation_of (Membership, P7)
annotation_of (Communication, P7)
annotation_of (Refund, P7)
annotation_of (Invoice, P8)
annotation_of (Delivery, P9)
annotation_of (ProductPlanning, P10)
annotation_of (PromotionPlanning, P10)
annotation_of (SupplierPolicy, P2)
annotation_of (ProductPlanning, P11)
annotation_of (StorePlanning, P11)
annotation_of (SupplierPolicy, P12)
annotation_of (StorePlanning, P3)
annotation_of (InventoryPlanning, P3)
annotation_of (StorePlanning, P13)
annotation_of (ProductPlanning, P13)
annotation_of (StorePlanning, P2)
annotation_of (OrderStatus, P9)
Continued on next page

Table 3.4 – continued from previous page

#annotation_of ($annotation, $process)
annotation_of (StaffPolicyDeveloping, P14)
annotation_of (SafetyPolicyExecute, P14)
annotation_of (InternshipProgramDeveloping, P14)

Table 3.5: Knowledge fact base (KFB) namely, fact_of($fact, $strategy) for facts and their relevant strategies used in Pyke.

#fact_of ($fact, $strategy)
fact_of (HappyCustomer, S1)
fact_of (LowRetailPrice, S1)
fact_of (LowProducingCost, S1)
fact_of (HappyCustomer, S2)
fact_of (LowRetailPrice, S2)
fact_of (GoodService, S2)
fact_of (SufficientQuantity, S2)
fact_of (ProductQuality, S2)
fact_of (Promotion, S2)
fact_of (ReliableSupplier, S3)
fact_of (ReliableSupplier, S4)
fact_of (HappyCustomer, S5)
fact_of (ProductQuality, S5)
fact_of (HappyCustomer, S6)
fact_of (SufficientQuantity, S6)
fact_of (ProductQuality, S6)
fact_of (ProductQuality, S7)
fact_of (SufficientQuantity, S7)
Continued on next page

Table 3.5 – continued from previous page

#fact_of ($fact, $strategy)
fact_of (LowWholesalePrice, S7)
fact_of (LowProducingCost, S7)
fact_of (ReliableSupplier, S7)
fact_of (HappyCustomer, S8)
fact_of (ProductQuality, S8)
fact_of (LowRetailPrice, S8)
fact_of (Promotion, S8)
fact_of (HappyCustomer, S9)
fact_of (OrderPolicy, S9)
fact_of (ConvenientStore, S9)
fact_of (StoreDecoration, S9)
fact_of (HappyCustomer, S10)
fact_of (OrderPolicy, S10)
fact_of (ConvenientStore, S10)
fact_of (StoreDecoration, S10)
fact_of (HappyCustomer, S11)
fact_of (OrderPolicy, S11)
fact_of (ConvenientStore, S11)
fact_of (ConvenientStore, S11)
fact_of (StoreDecoration, S11)
fact_of (HappyCustomer, S12)
fact_of (OrderPolicy, S12)
fact_of (ConvenientStore, S12)
fact_of (StoreDecoration, S12)
fact_of (HappyCustomer, S13)
fact_of (NewChannelsServices, S13)

Continued on next page

Table 3.5 – continued from previous page

#fact_of ($fact, $strategy)
fact_of (HappyCustomer, S14)
fact_of (NewChannelsServices, S14)
fact_of (HappyStaff, S15)
fact_of (StaffPolicy, S15)
fact_of (SafetyPolicy, S15)
fact_of (InternshipProgram, S15)
fact_of (HappyStaff, S16)
fact_of (StaffPolicy, S16)
fact_of (SafetyPolicy, S16)
fact_of (InternshipProgram, S16)
fact_of (HappyStaff, S17)
fact_of (StaffPolicy, S17)
fact_of (SafetyPolicy, S17)
fact_of (InternshipProgram, S17)
fact_of (HappyStaff, S18)
fact_of (StaffPolicy, S18)
fact_of (SafetyPolicy, S18)
fact_of (InternshipProgram, S18)
fact_of (HappyStaff, S19)
fact_of (StaffPolicy, S19)
fact_of (SafetyPolicy, S19)
fact_of (InternshipProgram, S19)
fact_of (HappyStaff, S20)
fact_of (StaffPolicy, S20)
fact_of (SafetyPolicy, S20)
fact_of (InternshipProgram, S20)

Continued on next page

Table 3.5 – continued from previous page

#fact_of ($fact, $strategy)
fact_of (ProductQuality, S21)
fact_of (SavingEnergy, S21)
fact_of (HappyCustomer, S22)
fact_of (ProductQuality, S22)
fact_of (LowRetailPrice, S22)
fact_of (Promotion, S22)

As stated in the previous section, a process is annotated and a strategy is linked with a fact manually. Hence, the annotations and facts in this case study are assumed and produced from the operational and planning knowledge of a chain. Using those annotations and facts, a list of rules are developed to establish the relationships between processes and their relevant strategies. For instance, in a rule R1: (FinalPrice $\wedge$ ProductPlanning $\wedge$ StorePlanning) $\Longrightarrow$ (HappyCustomer $\wedge$ LowRetailPrice $\wedge$ LowProductionCost), the right side of the rule is a consequent condition, i.e., the outcomes an enterprise aims to achieve, in this case increasing customer satisfaction, and reducing retail price and production cost. The left side of the rule is an antecedent condition, i.e., to reach all targets in the consequent condition, the relevant actions an enterprise must take. For example, the relevant actions in rule R1 are setting new low prices for products, modifying the plans for manufacturing products and re-organising stores to reduce the cost. These rules are shown in Table 3.6. We executed each rule using our proposed inference model for the knowledge base shown in Tables 3.1 and 3.2. After executing our inference model, the processes and their relevant strategies produced by the model are shown in Table 3.6. Consider the following example:

R1: (FinalPrice $\wedge$ ProductPlanning $\wedge$ StorePlanning) $\Longrightarrow$ (HappyCustomer $\wedge$ LowRetailPrice $\wedge$ LowProductionCost)

After executing our inference model as described in the previous section and Algorithm 3.1, the processes and their relevant strategies produced by the model are determined as follows using the Pyke program mentioned earlier:

- Processes: P1, P2, P3, P4, P5, P6, P10, P11, P13.
- Relevant strategies: S1, S2, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S22.

In this case, P1, P2, P3, P4, P5, P6, P10, P11 and P13 processes are related to S1, S2, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14 and S22 strategies that are bounded by R1.

For the validation, we checked all processes and their relevant strategies produced for each rule listed in Table 3.6. We found that our model produced all inferences for all rules accurately. This was done by considering the processes listed in Table 3.1 and finding the strategies relevant to those processes in Table 3.2. In the above example, as the antecedent condition of the rule relates to price, product and store, our Algorithm 3.1 correctly selects P1, P2, P3, P4, P5, P6, P10, P11 and P13 in Table 3.1 which are related to the criteria 'price, product and store'. The consequent condition of the rule identifies strategies listed in Table 3.2 that are relevant to the customer, retail price and production cost.

To achieve efficiency and competitive advantage, a business needs to keep their current good staff members and recruit new talent by executing new incentive and bonus strategies. To show how to incorporate new items or change the existing inference system, new annotations and a new strategy with its facts are added into Tables 3.1 and 3.2. These new items are highlighted in italics in the last grey cells of the tables. For instance, *SalaryPayment* and *RecruitmentPlanning* are added as new annotations to process P14. *StaffPolicy* and

IncentiveBonusPolicy are added as new facts to the new strategy '*More bonuses instead of salary increase* (S23)'. Then, another rule (R11) is defined as below:

R11: (StaffPolicyDeveloping $\wedge$ SalaryPayment $\wedge$ RecruitmentPlanning) $\Longrightarrow$ (HappyStaff $\wedge$ StaffPolicy $\wedge$ IncentiveBonusPolicy)

After executing our inference model, our algorithm selects the new process and strategies as below:

- Processes: P14
- Relevant strategies: S15, S16, S17, S18, S19, S20, S23

Table 3.6: Case study results: process-strategy alignment after execution of each rule using annotation and facts defined in Table 3.1 and 3.2

Rule No	Rule	Process-Strategy Relationship	
		Process	**Strategy**
R1	(FinalPrice $\wedge$ ProductPlanning $\wedge$ StorePlanning) $\Longrightarrow$ (HappyCustomer $\wedge$ LowRetailPrice $\wedge$ LowProductionCost)	P1 P2 P3 P4 P5 P6 P10 P11 P13	S1 S2 S5 S6 S7 S8 S9 S10 S11 S12 S13 S14 S22

Continued on next page

Table 3.6 – continued from previous page

Rule No	**Rule**	**Process-Strategy Relationship**	
		Process	**Strategy**
R2	((Feedback $\vee$ Membership $\vee$ Communication $\vee$ Refund $\vee$ Invoice $\vee$ Delivery) $\wedge$ ProductPlanning $\wedge$ PromotionPlanning $\wedge$ FinalPrice $\wedge$ StorePlanning) $\Longrightarrow$ (HappyCustomer $\wedge$ LowRetailPrice $\wedge$ GoodService $\wedge$ SufficientQuantity $\wedge$ ProductQuality $\wedge$ Promotion)	P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P13	S1 S2 S5 S6 S7 S8 S9 S10 S11 S12 S13 S14 S21 S22
R3	(SupplierPolicy $\wedge$ StorePlanning) $\Longrightarrow$ Reliable	P2 P6 P11 P12 P13	S3 S4 S7
		P2	S1 S2 S5 S6 S7

Continued on next page

Table 3.6 – continued from previous page

Rule No	Rule	Process-Strategy Relationship	
		Process	**Strategy**
R4	(ProductPlanning ∧ PromotionPlanning ∧ StorePlanning) ⟹ HappyCustomer ∧ SufficientQuantity ∧ ProductQuality	P3 P4 P5 P6 P10 P11 P13	S8 S9 S10 S11 S12 S13 S14 S21 S22
R5	(SupplierPolicy ∧ StorePlanning ∧ ProductPlanning) ⟹ (ProductQuality ∧ SufficientQuantity ∧ LowWholesalePrice ∧ LowProductionCost ∧ ReliableSupplier	P2 P3 P4 P5 P6 P10 P11 P12 P13	S1 S2 S3 S4 S5 S6 S7 S8 S21 S22
R6	(InventoryPlanning ∧ ProductPlanning ∧ PromotionPlanning ∧	P2 P3 P4	S1 S2 S5 S6 S7 S8

Continued on next page

Table 3.6 – continued from previous page

Rule No	Rule	Process-Strategy Relationship	
		Process	**Strategy**
	StorePlanning) $\Longrightarrow$ (HappyCustomer $\wedge$ ProductQuality $\wedge$ LowRetailPrice $\wedge$ Promotion	P5 P6 P10 P11 P13	S9 S10 S11 S12 S13 S14 S21 S22
R7	(Feedback $\vee$ Membership $\vee$ Communication $\vee$ Refund $\vee$ Invoice $\vee$ Delivery) $\wedge$ ProductPlanning $\wedge$ PromotionPlanning $\wedge$ StorePlanning $\wedge$ SupplierPolicy $\Longrightarrow$ (HappyCustomer $\wedge$ OrderPolicy $\wedge$ ConvenientStore $\wedge$ StoreDecoration	P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13	S1 S2 S5 S6 S8 S9 S10 S11 S12 S13 S14 S22
	((Feedback $\vee$ Membership $\vee$ Communication $\vee$ Refund $\vee$ Invoice $\vee$	P2 P3 P4	S1 S2 S5 S6

Continued on next page

Table 3.6 – continued from previous page

Rule No	Rule	Process-Strategy Relationship	
		Process	Strategy
R8	OrderStatus $\vee$ Delivery) $\wedge$ ProductPlanning $\wedge$ StorePlanning) $\implies$ (HappyCustomer $\wedge$ NewChannelsServices	P5 P6 P7 P8 P9 P10 P11 P13	S8 S9 S10 S11 S12 S13 S14 S22
R9	StaffPolicyDeveloping $\wedge$ SafetyPolicyExecute $\wedge$ InternshipProgramDeveloping $\implies$ HappyStaff $\wedge$ StaffPolicy $\wedge$ SafetyPolicy $\wedge$ InternshipProgram	P14	S15 S16 S17 S18 S19 S20
R10	StorePlanning $\wedge$ ProductPlanning $\implies$ ProductQuality $\wedge$ SavingEnergy	P2 P3 P5 P6 P10 P11 P12 P13	S2 S5 S6 S7 S8 S21 S22

3.3 Conclusions

The existing studies of process-strategy alignment have shortcomings, as discussed in Chapter 2. Specifically, the methods proposed for the process-strategy relationship are not sufficiently flexible for application in diverse systems; rather, each was designed only for a single specific system. Due to the complexity of business processes and the ambiguity in using natural language for describing business strategies, it is difficult to determine process-strategy relationships between business processes and strategies. This also leads to great uncertainty in discovering the relationship between business processes and their relevant strategies.

In this chapter, an inference model has been proposed to deal with these issues. The aim of the inference model based on business rules is to establish the relationship between processes and their relevant strategies via their annotations and facts. By considering annotations and facts from the rules defined by business experts, automatic determination can fill the gap in the process-strategy relationship. The allocation of annotations to each business process reduces the complexity of business processes, while business strategy standardization by using facts eases the ambiguity of business strategies. This model has also proven its ability to cover all important relationships between business processes and their relevant strategies by inference-based on defined rules. The model has been implemented in the Python environment using the Pyke expert engine. The results produced by our proposed model considering a business case show that our model can accurately and automatically identify the relationships between business processes and their corresponding strategies.

The outcome of this chapter has high applicability in many cutting edge applications (e.g., drones, driverless cars, industrial robotic systems). The model proposed here is currently applied to the specific scenario 'Retail', but with minor modifications, it can be readily applied to automation in any deployment environment where uncertainty exists because of dynamic environmental changes.

As indicated in Chapter 1, one of the main objectives of this research project is to develop the significance level of a query considering the business processes and strategies of a business organisation. How the significance level of a query can be derived considering the process-strategy relationship of a business organisation developed in this chapter is presented in the next chapter.

Chapter 4

Significance Level of Text-Based Big Data Queries

For a business organisation, the importance of the dynamic scaling of semantic information to be applied in both big data collection and analytics has been highlighted in Chapter 1 and presented as a research objective (Objective 1). The significance level of a query can make a trade-off between query response delay and the extent of data collection and analysis, and thus can provide a solution for Objective 1. This motivates us to concentrate on determining the significance level of a query based on its importance to an enterprise system. To our knowledge, no such approach is available in the literature. To address this research gap, we aim to determine the significance level of a query in order to prioritise the query with relevance to a business organisation. In this chapter, we propose two solutions as follows:

(i) In the first solution, we determine the significance level of a query based on the semantic similarity between a query and the processes of a business organisation and the importance of a business process to an organisation. The core business processes are given higher importance than non-core business processes. We name our proposed approach to this solution the 'Significance level of a query based on business processes'.

(ii) In the second solution, to link the significance level with business processes and business strategies, the significance level of a query is calculated based on the consideration of both business processes and strategies. This is done by exploiting process contributions and strategy priorities. We refer to this approach as the 'Significance level of a query considering business process and business strategy' throughout the chapter.

In addition, we present a business case study for both solutions to show the difference between them. The structure of this chapter is as follows: the first solution is presented in Section 4.1 and the second solution is presented in Section 4.2. The results of the case study for each solution are also described in each section and Section 4.3 concludes this chapter.

4.1 Significance Level of Query Considering BP (SLQBP)

4.1.1 Description of SLQBP approach

One of the most important uses of a query in business intelligence is to collect and analyse data in order to discover some clues or gain deep insights into the data that can help achieve efficiency gains or a competitive advantage over other similar organisations. The business processes of an enterprise system are also used to obtain efficiency gains and competitive advantage over others. This implies that querying the data must meet the main purpose of business processes. However, for an enterprise organisation, not all business processes are equally important. Some business processes are considered core business processes, while others are regarded as non-core business processes.

To successfully achieve the business goals and combat computational complexity in a tactical way, the result of querying data should reflect the core business processes more extensively than others. This motivates us to calculate the significance of a query based on the importance of the business processes of an organisation. This calculation involves two main items (i) which

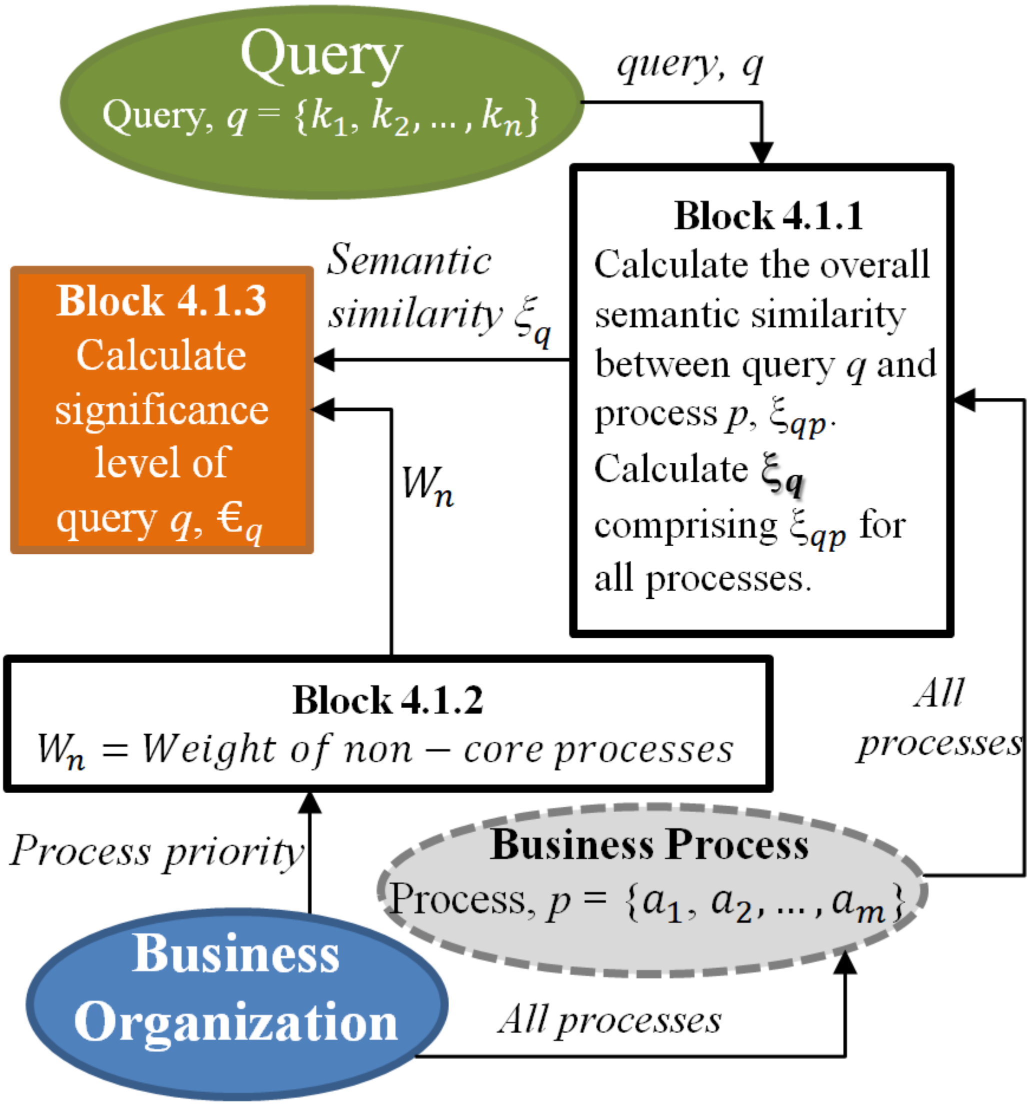

Figure 4.1: Overall process of our proposed approach (Significance Level of a Query Considering Business Processes) to determine the significance level of a query

queries are related to which business processes and to what extent? and (ii) which processes are how important to a business organisation? For the former, we consider the semantic similarity between the keywords of a query and the annotations of a business process. This is because the keywords of a query represent the purpose of the query, while the annotations of the business process denote the activities and the other aspects of a business process. For the latter, following the widely used and well-adopted approach in business organisations, the importance of a business process is attributed according to whether a business process is a core or non-core.

The concept of semantic information is also valuable to model business processes. Business processes that are semantically annotated show enrichment in their process descriptions. A set of carefully selected semantic annotations for a business process can not only reduce the ambiguity in representing the logically connected activities of that business process for fulfilling a certain business goal, but also provide business analysts with a better understanding of the business processes. In addition, semantic annotations can also make a relationship between the business processes and the other characteristics of an organisation. Therefore, the semantic annotation scheme allows us to link a query, business processes and business strategies semantically by avoiding mismatched and unstructured knowledge representation in a business process model [160,161]. This has motivated us to calculate a similarity score between a query and a business process by exploiting the semantic similarity scores between the keywords of a query and the annotations of a business process.

Suppose, query q consists of n keywords $\{k_1, k_2, \cdots, k_n\}$ and process p contains m annotations $\{a_1, a_2, \cdots, a_m\}$ where k_i is i^{th} keyword of the query q and a_j is the j^{th} annotation of process p. The similarity score ξ_{qp} between a process p and a query q is calculated using method 2.1 described in Chapter 2:

$$\xi_{qp} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\xi(k_i, a_j)}{n \times m} \tag{4.1}$$

We then form a vector ξ_q comprising all ξ_{qp}.

Assuming the significance level of a query varies non-linearly with the semantic similarity value between a query and a business process, we define the significance level of query q for a business organisation as:

$$€_q = \frac{e^{\bar{\xi}_q} + e^{\tilde{\xi}_q \times w_n}}{e + e^{w_n}} \tag{4.2}$$

where, $\bar{\xi}_q$ and $\tilde{\xi}_q$ are the median value of ξ_q for the core and non-core business processes, respectively, and $W_n \in [0, 0.5]$ represents the weight for non-core processes selected by a business organisation according to the requirements of the enterprise system. The overall process of our proposed system is illustrated in Figure 4.1.

In Figure 4.1, all the business processes of an enterprise organisation and a query of an end-user are fed to Block 4.1.1. Block 4.1.1 calculates the overall semantic similarity ξ_{qp} between query q and process p as shown in (4.1). In this way, the semantic similarities between the query and all business processes are calculated and finally ξ_q is obtained. ξ_q is provided as an input to Block 4.1.3. Block 4.1.2 captures the weights W_n for non-core business processes from the business organisation and supplies this to Block 4.1.3. Utilizing W_n and ξ_q, Block 4.1.3 calculates the significance level of query q, $€_q$ by applying (4.2).

4.1.2 A case study of SLQBP

The proposed approach is implemented using the Python programming language. It was validated by considering a business case of a retail company. Table 4.1 shows that the business organisation has 14 business processes. The core business processes are nominated depending on the business strategies. These processes indicate the aspect on which an enterprise organisation should focus with higher priority in order to reach their goals. We have identified the following five business processes as the core business processes as these processes are highly related to customers and human resources:

P1. Customer service (after sales service)
P7. Marketing research
P8. Product design/development
P10. Business planning
P14. Personnel management

The remaining nine processes listed below are the supplementary (non-core) processes:

P2. Sales/order entry (selling and entering orders)
P3. Invoicing/billing (generation and mailing of invoices/bills)
P4. Purchasing (ordering from suppliers)
P5. Advertising/promotion
P6. Pricing
P9. Distribution (transporting goods to market)
P11. Inventory management (keeping inventories at planned levels)
P12. Quality management (measuring, monitoring, and taking action to maintain quality)
P13. Production scheduling/planning (for manufacturing requirements)

Table 4.1: List of business processes

Business process	**Abbreviated name**
Customer service (after sales service)	P1
Sales/order entry (selling and entering orders)	P2

Continued on next page

Table 4.1 – continued from previous page

Business process	**Abbreviated name**
Invoicing/billing (generation and mailing of invoices/bills)	P3
Purchasing (ordering from suppliers)	P4
Advertising/ promotion	P5
Pricing	P6
Marketing research	P7
Product design/development	P8
Distribution (transporting goods to market)	P9
Business planning	P10
Inventory management (keeping inventories at planned levels)	P11
Quality management (measuring, monitoring and taking action to maintain quality)	P12
Production scheduling/planning (for manufacturing requirements)	P13
Personnel management	P14

Table 4.2: List of big data text-based queries

Big data query	**Abbreviated name**
How many existing customers referred new customers to your company?	Q1
Which product information pages have the most views?	Q2

Continued on next page

Table 4.2 – continued from previous page

Big data query	Abbreviated name
What is the percentage of abandoned shopping carts?	Q3
What is the total paid revenue vs. outstanding orders?	Q4
What is the total number of returned/ cancelled orders and revenue loss from them?	Q5
What customers said about the promotion last month	Q6
Are customers happy with the quality of fruit and vegetables	Q7
Do customers like shopping online and how frequently do they do it	Q8
What is the current shopping trend?	Q9
Satisfaction of customers with frozen fruits	Q10
Determine customer satisfaction levels with fruit and vegetables	Q11
What do customer think about fruit and vegetable	Q12
Are staff happy about our training program?	Q13
Racism and terrorism in recent years	Q14
The trend of same-sex marriage in the world	Q15

Continued on next page

Table 4.2 – continued from previous page

Big data query	**Abbreviated name**
Should gym and diet always come together?	Q16
Are dogs better than cats?	Q17
Obesity epidemic in school-aged children	Q18

To calculate the significance level for a number of different queries, we used 18 different big data text-based queries which are generally used in an enterprise system of a retail company. These big data queries are listed in Table 4.2 where entire sets of queries comprising of keywords and their abbreviations are also presented. For example, the abbreviated name of query number 5 is Q5.

We used weight $W_n = 0.5$ for non-core business processes, $\alpha = 0.2$ and $\beta = 0.45$ as suggested in [159]. The semantic similarity values between 18 queries (Q1 to Q18) and 14 business processes (P1 to P14) produced by our proposed approach using (4.1) are shown in Table 4.3.

Table 4.3 shows the similarity scores between the queries and business processes calculated using the method presented in Section 4.1. Our aim is to show how each business process is related to individual queries. Table 4.3 clearly shows that query Q1 has highest semantic similarity score (1.0) with business processes P6, P7 and P8. This means that processes P6, P7 and P8 are the most relevant to query Q1. The lowest semantic similarity score (0.15) for query Q8 is with process P9. Therefore, process P9 is the least relevant to query Q8.

Table 4.3: Similarity scores between queries and processes

Pl / Qj	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**
Q1	0.86	0.26	1.00	0.20	0.74	1.00	1.00	1.00	0.19	0.38	0.32	0.30	0.17	0.45
Q2	0.34	0.69	0.46	0.57	0.57	0.48	0.55	0.55	0.29	0.56	0.53	0.61	0.56	0.43
Q3	0.29	0.42	0.28	0.73	0.50	0.33	0.46	0.46	0.31	0.46	0.42	0.56	0.40	0.60
Q4	0.25	0.41	0.34	0.49	0.45	0.29	0.44	0.44	0.17	0.41	0.30	0.45	0.37	0.51
Q5	0.42	0.56	0.39	0.56	0.56	0.43	0.47	0.45	0.18	0.56	0.39	0.56	0.47	0.51
Q6	0.60	0.39	0.46	0.40	0.57	0.56	0.43	0.43	0.25	0.37	0.41	0.40	0.29	0.57
Q7	0.43	0.41	0.43	0.47	0.62	0.39	0.56	0.56	0.22	0.42	0.28	0.49	0.40	0.38
Q8	0.36	0.49	0.41	0.43	0.54	0.35	0.52	0.52	0.15	0.67	0.21	0.42	0.40	0.38
Q9	0.36	0.44	0.44	0.44	0.52	0.37	0.71	0.62	0.26	0.44	0.32	0.40	0.38	0.36
Q10	0.52	0.45	0.47	0.43	0.64	0.49	0.59	0.59	0.28	0.44	0.28	0.45	0.41	0.34
Q11	0.45	0.42	0.41	0.46	0.62	0.46	0.54	0.54	0.31	0.41	0.30	0.48	0.37	0.42
Q12	0.40	0.42	0.46	0.40	0.63	0.45	0.57	0.57	0.22	0.45	0.29	0.43	0.42	0.28
Q13	0.47	0.45	0.44	0.50	0.53	0.37	0.39	0.39	0.23	0.53	0.41	0.38	0.41	0.70
Q14	0.25	0.34	0.28	0.37	0.31	0.21	0.34	0.28	0.23	0.38	0.36	0.42	0.28	0.40
Q15	0.39	0.49	0.26	0.42	0.46	0.35	0.53	0.50	0.21	0.40	0.35	0.45	0.30	0.46
Q16	0.21	0.29	0.24	0.34	0.45	0.15	0.29	0.29	0.21	0.32	0.28	0.29	0.29	0.29
Q17	0.37	0.27	0.37	0.36	0.42	0.28	0.33	0.33	0.23	0.37	0.26	0.38	0.27	0.45
Q18	0.32	0.50	0.25	0.43	0.39	0.38	0.30	0.28	0.24	0.33	0.36	0.48	0.27	0.49

Pl = Process, Qj = Query

As all business processes were divided into two groups by their own importance to the business organisation (core or non-core), the overall significance level of each query is then calculated based on these weights. The significance levels of queries Q1 to Q18 calculated by our proposed approach are presented in Figure 4.2.

In Figure 4.2, we can see that the highest significance level (0.8060) is for Q1: How many existing customers referred new customers to your company? This is mainly because Q1 has the highest value of the significance levels with more core-business processes (P1, P7, P8, P10, P14) than the other queries. This vindicates the reasons why Q1 is the most important query of the enterprise system considered in this case study.

Since Q1 is the most important query, both data collection and analytic techniques should exhaustively consider the semantic information of its keywords to obtain deep insights into the data. The consideration of semantic information about a query also reduces unnecessary data collection and thus eventually decreases the amount of data to be analysed, while retaining all relevant data. As a consequence, this speeds up the computational process of the data analytic techniques compared to the case where all data are collected without considering semantic information. This increased will bring much benefit to enterprise information systems as today each business needs to handle enormous amounts of growing data. In addition, in this way data analytics tools at a later stage will see less non-relevant data and therefore, better and more accurate analytical outcomes are expected.

As mentioned in Chapter 2, semantic information helps in capturing the deep insights of data by extending the meaning of data more deeply via conceptual and lexical relations. However, this leads to time consumption in data processing. Therefore, considering semantic information should only concentrate on the queries which are important to an enterprise organisation. As the keywords of query Q16 'should gym and diet always come together?' are not relevant to the scenario of a retail company, this query has the lowest significance level (0.57). This means that Q16 is the least important query to the company compared with the other queries. Hence, by not considering the semantic information, the amount of time to process this query could be reduced and this could be a stepping stone to the enhanced the efficiency of the overall system.

4.2 Significance Level of Query Considering BP and BS (SLQBPS)

The significance Level of a Query Considering Business Processes, introduced in Section 4.1, has proven that by considering the significance level of a query, the amount of time spent on less important queries can be reduced. For capturing deep insights, the data processing system can focus more on the queries that have higher significance levels. However, in this approach, only business processes are considered by manually selecting weights for core and non-core business processes. Furthermore, since the business strategies are not utilised at all, the significance level of a query calculated by this approach does not fully cover the objectives of a business organisation and hence cannot capture deep insights into data relevant to fulfilling the goals of an organisation.

As the significance level of a query must express the importance of that query to a business organisation, it must be calculated based on the consideration of the business strategic direction, which requires the consideration of both business processes and strategies. Solution 2 in this sub-section proposes

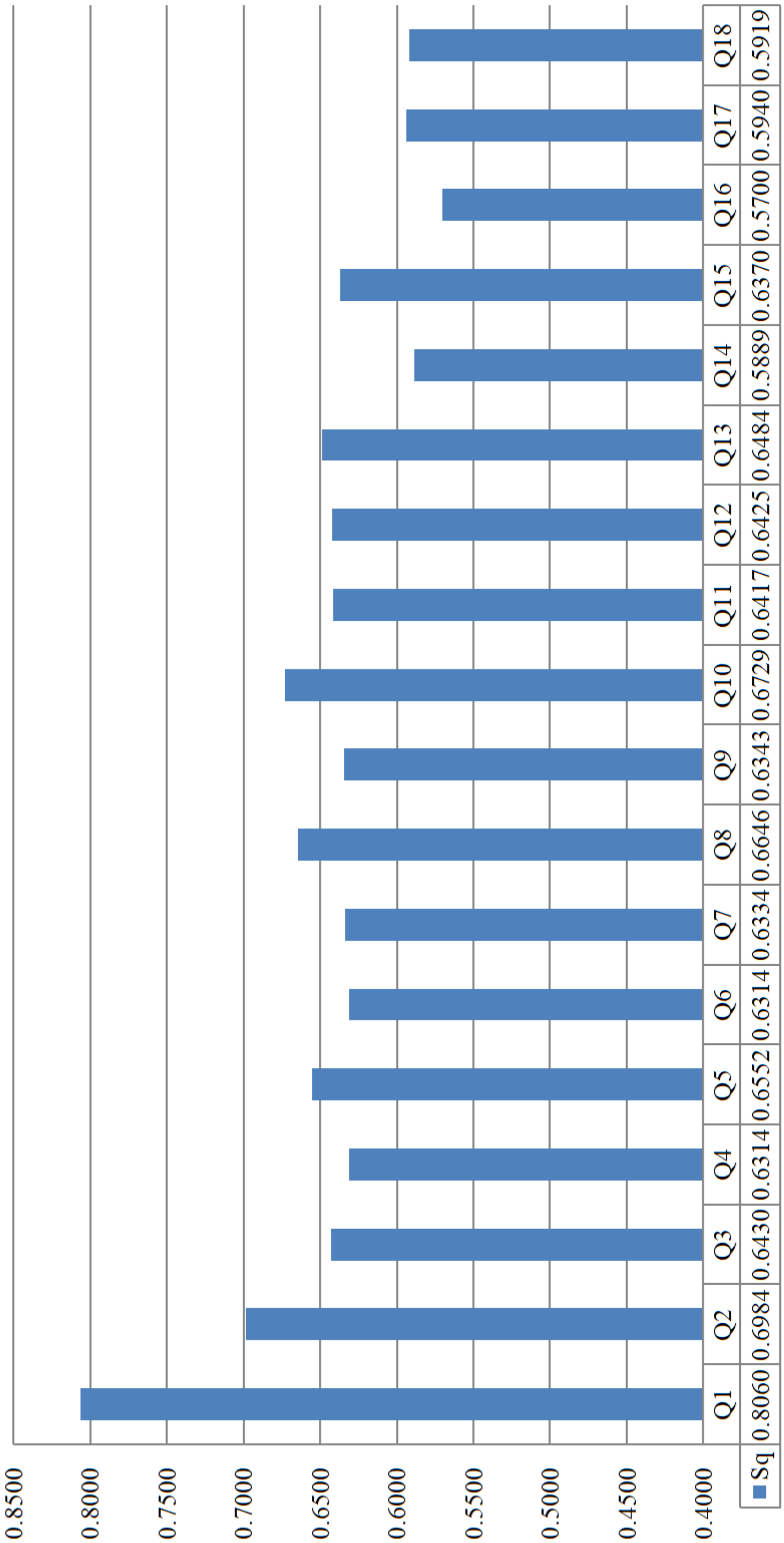

Figure 4.2: Significance levels of queries for a business organisation

an approach where the significance level of a query is determined by exploiting process contributions and strategy priorities.

As no studies to date have worked on calculating the significance level of a query based on the aspect of business strategy, for the first time, we propose a method that determines the significance level of a query reflecting the business objectives. The contributions of this research are highlighted as below:

- We determine the significance level of a query considering both business processes and organisational strategies.
- The business process contributions are calculated based on the extent of the contribution of a process to business strategies, the number of strategies to which a process contributes, and the priority score of each strategy for the business organisation.
- The results produced by the proposed approach in our business case study have shown that queries related to more important business processes and higher priority strategies, have higher significance levels.

The technique used in the proposed approach is presented below.

4.2.1 Description of the SLQBPS technique

As stated previously, the significance level of a query needs to reflect the importance of that query for a business organisation. Queries are normally very short and a business organisation is a combination of many activities and transactions. These raise an important question 'how to find a link between a query and a business organisation'. To answer this question, we have decided to choose business process and strategy as the main representation for a business organisation. This is because business processes and strategies are the two key entities that always play the main role in business organisation [141]. As mentioned in Section 4.1, business processes represent all the activities and transactions in business organisations. Business strategies play vital roles in business development including earning more revenue, achieving strategic advantage and its expansion. A significance level of a query is thereby determined based on the relationship between query, process and strategy as seen in Figure 4.3.

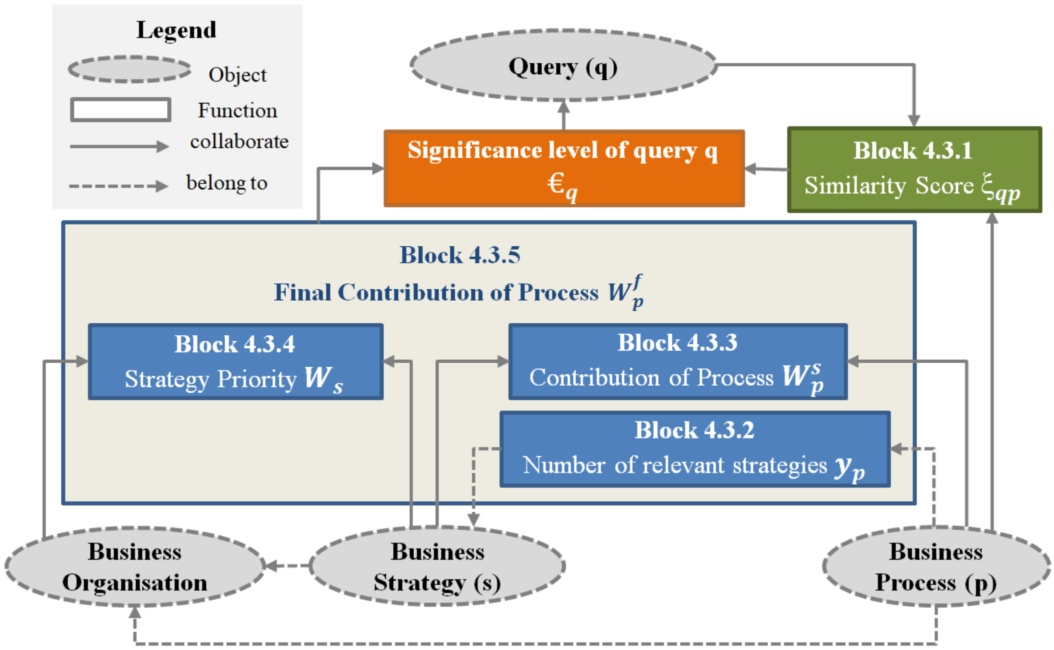

Figure 4.3: The overall process of proposed approach to determine the significance level of a query

In Section 4.1, to determine the significance level of a query, the contribution of a process is intuitively assigned as two values, i.e., 0.5 or 1. The value 1 is for core business processes, while 0.5 for non-core business processes. Dividing processes into two groups (i) core and (ii) non-core does not fully reflect the strategic direction of a business organisation. Because some processes may have either very high or very low effect on a business organisation, while the others may incur difficult levels of effect ranging between the two extreme values. This indicates that the contribution of a process should be a continuous value in [0, 1]. On the other hand, the business strategies that describe business tactics to achieve the business objectives and goals have not been taken into account in determining the significance level of query. Therefore, this demands the introduction of a new approach that is able to calculate the significance level of a query reflecting a process contribution more accurately and integrating the business strategies.

To embed the more accurate reflection of a business objective in a big data query, we aim to introduce an approach to calculate the significance level of a query based on the contribution weight of a process and the priority level of a strategy. Note that, a contribution weight refers to how much impact a process p has on the strategy satisfaction [162]. For example, Table 4.4 shows strategy S3 has five contributing processes P2, P4, P5, P6 and P12, whose contributions are 90%, 70%, 70%, 50% and 80%, respectively.

The term "strategy priority" refers to the task of ranking strategic objectives based on their importance in business organisation. By clarifying their roles and urgency, a business organisation has a better view of what to do first, what need to be focused more than the others [163, 164].

To consider process contributions to strategies and a strategy priority level in determining the significance level of a query, firstly, we need to find the process-strategy relationship. This can be done using the model introduced in Chapter 3 where the link between a strategy and a process is built by adopting a rule-based inference model.

Besides the importance of realising which strategies a process contributes to, secondly, we then need to indicate how successful a strategy will be when

Table 4.4: Contribution weight of processes

Ps / Si	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
S1	90%	90%	20%	90%	90%	50%	0%	0%	50%	65%	10%	65%	55%	0%
S2	50%	90%	45%	60%	90%	90%	70%	55%	50%	70%	20%	70%	70%	55%
S3	0%	90%	0%	70%	70%	50%	0%	0%	0%	0%	0%	80%	0%	0%
S4	0%	90%	0%	60%	70%	50%	0%	0%	0%	0%	0%	80%	0%	0%
S5	0%	90%	80%	75%	75%	80%	0%	0%	0%	60%	0%	80%	70%	0%
S6	0%	90%	90%	85%	85%	90%	0%	0%	0%	0%	80%	80%	70%	0%
S7	0%	90%	0%	70%	70%	50%	0%	0%	0%	0%	0%	80%	0%	0%
S8	0%	90%	0%	85%	90%	75%	0%	0%	0%	75%	0%	85%	90%	0%
S9	0%	90%	35%	10%	10%	10%	70%	50%	50%	65%	65%	0%	85%	40%
S10	0%	90%	70%	65%	65%	20%	85%	75%	80%	75%	75%	0%	75%	0%
S11	0%	90%	10%	50%	50%	75%	70%	50%	60%	70%	75%	0%	85%	0%
S12	0%	90%	10%	65%	65%	45%	0%	0%	0%	75%	65%	0%	70%	0%
S13	0%	90%	40%	85%	85%	65%	70%	40%	70%	80%	70%	0%	75%	0%
S14	0%	90%	70%	90%	90%	35%	70%	50%	0%	80%	75%	0%	80%	50%
S15	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%
S16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%
S17	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%
S18	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%
S19	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	80%
S20	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	80%
S21	0%	90%	90%	50%	50%	55%	0%	0%	0%	0%	55%	0%	0%	0%
S22	70%	90%	75%	70%	90%	70%	0%	0%	0%	80%	55%	80%	85%	0%
S23	0%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%

Ps = Process, Si = Strategy

its related process is completely executed. This is called a contribution weight of a process for a specific strategy [162]. These contribution weights can be determined by the chief information officer or business analyst or the manager of a business organisation.

Thirdly, we need to consider how to determine the strategy priorities. A strategy priority represents the extent of importance and urgency of a strategy. This means the more the importance and urgency are, the higher is the priority of a strategy. By clarifying the strategy priority, a business organisation has a better view of what to do first and what needs to be focused on [163, 164].

In this approach, to determine the significance level of a query, five main values depicted in Figure 4.3 need to be calculated: (i) the similarity scores between process p and query q, ξ_{qp} (**Block 4.3.1**), (ii) the number of strategies y_p in which process p contributes to (**Block 4.3.2**), (iii) the contribution of process p to strategy s, W_p^s (**Block 4.3.3**), (iv) the strategy priority of strategy s, W_s (**Block 4.3.4**), and (v) the final contribution of process p, W_p^f using y_p, W_p^s and W_s (**Block 4.3.5**). Note, as described above, Block 4.3.5 represents the function for the calculation of final contribution process W_p^f. In Figure 4.3, this block also shows that the calculation of W_p^f requires the execution of the functions represented by Blocks 4.3.2, 4.3.3 and 4.3.4.

The similarity score ξ_{qp} between a process p and a query q is calculated as described using 4.1 in Solution 4.1. The results of similarity scores between the queries and business processes are as shown in Table 4.3.

As mentioned before, the contribution of process p to strategy s, W_p^s can be assigned by the chief information officer or business analyst. The strategy priority W_s of the strategy s can be determined based on a business strategy prioritization tool [165].

The final contribution of a process is then calculated as follows:

(i) The contribution of a process should be determined by using a utility function that follows the diminishing property of microeconomic model [166], i.e., the contribution of a process should follow a logarithmic function. This means that similar to logarithmic curve, it should increase with increasing number of associated strategies up to a certain limit and after that the increment in contribution should slow down. Therefore, we can define the contribution of process p, W'_p as,

$$W'_p = \frac{log(y_p^c + 1)}{log(z^c + 1)} \tag{4.3}$$

where, z is the total number of strategies considered, $0 \leq W'_p \leq 1$ for $y_p \leq z$ and c represents the sensitivity of the process contribution with respect to the number of associated strategies. The process contribution (W'_p) derived in (4.3) was plotted over the number of associated strategies (y_p) for different values of c=0.2, 0.5, 1.0, 2.0 and z=23, and is shown in Figure 4.4. Figure 4.4 shows for a constant value of z, the higher the value of c, the higher sensitivity of the process contribution on the number of business strategies in which a process contributes to. Therefore, for a particular business process, if it requires to quickly vary the process contribution over the number of associated strategies, the higher value of c needs to be used.

(ii) Considering the contribution of a process to all strategies, and the values of the strategies' priority, we can define the following aggregation function to calculate the contribution of a process p, W_p^c using W_p^s and W_s as,

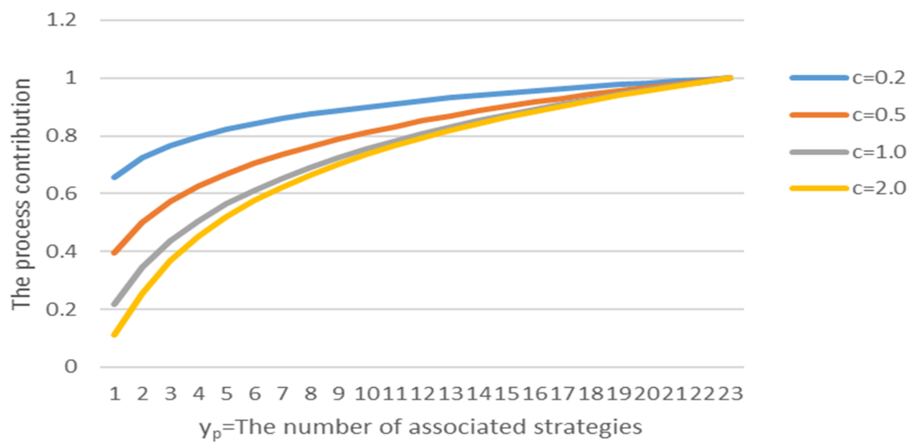

Figure 4.4: The process contribution vs the number of associated strategies for different values of c=0.2, 0.5, 1.0, 2.0 and z=23

$$W_p^c = \frac{\sum_{s=1}^{z}(W_p^s \times W_s)}{z} \tag{4.4}$$

Equation (4.4) exhibits a consensus view that the higher the values of W_p^s and W_s, the higher value of W_p^c is.

Assuming the contribution of a process to the extent of covering the number of strategies, and the amount of contribution to all strategies and their priority levels as equally important, we can formulate an equation to determine the final contribution of process p in terms of a combination of W_p' and W_p^c defined in (4.3) and (4.4), respectively in the following way,

$$W_p^f = \frac{W_p' \times W_p^c}{\max(W_p^f)} \tag{4.5}$$

where, $max(W_p^f)$ is the maximum value of all W_p^f and used to normalise the value in [0,1].

Finally, similar to (4.3), assuming the significance level of a query varies non-linearly following a logarithmic curve with its similarity score with a process and the process contributions, the significance level of a query q can be determined using the similarity scored ξ_{qp} defined in (4.1) and the final contribution W_p^f derived in (4.5) as,

$$€_q^s = \frac{log(\sum_{p=1}^{t}(W_p^f \times \xi_{qp}) + 1)}{log(t+1)} \tag{4.6}$$

where, t is the total number of business processes.

4.2.2 A case study for SLQBPS

The proposed approach was implemented using Pyke package in Python programming language for establishing the relationship between a process and a strategy. The contribution weights of the processes were intuitively assigned and the strategy priorities shown in Table 4.7 were determined by using the strategy prioritization tool [165]. This case study also uses the same set of processes and queries as shown in Tables 4.1 and 4.2, respectively. The list of strategies is presented in Table 4.5.

Table 4.5: List of business strategies

Business Strategy	**Abbreviated name**
Lowering the cost of planning	S1
Maintaining great quality of own brands	S2
Building strong, collaborative partnership with suppliers	S3

Continued on next page

Table 4.5 – continued from previous page

Business Strategy	**Abbreviated name**
Long-term partnership	S4
Offering fresh product to customers	S5
Improve the quality and availability of fresh food	S6
Long standing suppliers	S7
Developing an innovative and unique product	S8
Improving store network across all own brands	S9
Making shopping easier and simpler for customers	S10
Creating bigger, brighter stores with new features	S11
New-look liquor-land stores	S12
Opening first online standalone store	S13
Expending into new channels and services	S14
Bringing together diverse backgrounds	S15
Launching first accessibility action plan	S16
Increasing the number of women in leadership	S17
Supporting indigenous team members	S18
Maintaining a safe workplace	S19
Maintaining graduate program	S20
Natural refrigerants	S21
Growing exclusive brands	S22
More bonuses instead of salary increase	S23

According to the tool [165] mentioned earlier, the strategy priority is determined by following three main criteria - (i) strategic fit, (ii) economic impact and (iii) feasibility. The elements of each criteria were weighted depending on their importance. The element of each criteria for a specific strategy was then

Table 4.6: An example for strategy priority of S1

Criteria	Element	Weight	Rank for S1
Strategic fit	Alignment with company goals	15%	9
	Market positioning	15%	9
	Core capabilities	10%	7
Economic impact	Revenue potential	10%	7
	Profitability & margin	15%	6
	Growth potential	15%	8
Feasibility	Technical risk	10%	1
	Resources - Financial	5%	8
	Resources - People	5%	7
Strategy priority for S1			7.1

ranked between 1 to 10. The more important an element is, the higher rank it has. As a representative example, the calculated rank of each element of strategy S1 including S1's priority value (7.1) is presented in Table 4.6. Finally, the calculated priority value of each strategy is listed in Table 4.7. The results show S2 having priority value 7.7 is the highest priority, while S18 with priority value 2.8 is the lowest.

The similarity score ξ_{qp} between each query and each process was calculated using (4.1). These scores are as shown in Table 4.3 that describe the highest ($0.9994 \approx 1.00$) similarity scores were obtained between Q1 and four processes P3, P6, P7, P8 and the lowest ($0.1505 \approx 0.15$) similarity score was between Q8 and P9.

According to a rule-based inference model introduced in Chapter 3, we have applied this model to our scenario and found that process P1 is associated with the strategies S1, S2 and S22. Similarly, the relationship of other processes with their relevant strategies was discovered using the same knowledge base. Then the contribution weights of all processes for each strategy were intuitively assigned and are given as described in Table 4.4. For example, process P1 has the highest contribution weight on strategy S1. This means after P1 is completely executed, strategy S1 is expected 90% chance of success. For those places where the contribution weights are 0% such as the contribution weights of P1 on the strategies S3 to S21 and S23, this means there is no rule in knowledge base that reflects a link between P1 and these strategies or there is no effect of P1 on these strategies. Another example, strategy S7 has five contributing processes P2, P4, P5, P6 and P12, whose contributions are 90%, 70%, 70%, 50% and 80%, respectively. This means if process P2 is completely executed, there is a 90% chance that strategy S7 is achieved. Similarly, if P4, P5, P6 and P12 are completely executed, there is a 70%, 70%, 50% or 80% of chance for strategy S7 to be achieved.

Table 4.7: Priority score of strategies

Strategy	Priority	Strategy	Priority
S1	7.1	**S13**	6.9
S2	7.7	**S14**	5.4
S3	4.8	**S15**	3.9
S4	5.4	**S16**	3.2
S5	7.1	**S17**	3.5
S6	7.4	**S18**	2.8
S7	5	**S19**	5.3
S8	5.4	**S20**	3.1
S9	4.5	**S21**	5.7
S10	7.8	**S22**	6.1
S11	4.9	**S23**	4.8
S12	4.8		

Figure 4.5 shows the significance level of all queries produced by our proposed approach using (4.6). To show the impact of c used in (4.3) on the significance level of all queries, we used three different values of c such as 0.5, 1 and 2 keeping the value of other parameters as the same. Figure 4.5 shows the lower the value of c, the higher the significance level of all queries is. This is because for a constant number of total strategies (e.g., 23 strategies have been used in this case study), the lower the value of c, the lower the effect of the number of strategies associated with a business process (y_p) is. Therefore, c provides a way to make a trade-off between W_p' and W_p^c to a certain extent. For the sake of clarity, in the remainder of this paper, we will discuss the results only for c = 1. The significance levels with c=1 are between 0.384 and 0.547. The results shows has the highest significance level (0.547). This is because, query Q2 has the highest similarity score 0.6864 with process P2. The contribution weights of process P2 with strategies S1 to S14 and S21 to S23 are 90%. The priority of strategies S1, S2, S5, S6, S10 and S13 are in the top of highest priority scores.

On the other hand, query Q16 has the lowest significance level as 0.384. This is mainly because this query has the highest similarity score 0.4519 with P5. Although, P5 also has contribution with similar group of strategies like P2 except S23 but the contribution weights of P5 are much lower than those of P2.

However, the significance level of the same queries computed by Significance Level of a Query Considering Business Processes presented in Section 4.1 without considering the contribution of a process to each strategy and the strategy priority is in [0.570, 0.806]. If we compare this range values with the values [0.384, 0.547] that are determined by Significance Level of a Query Considering Business Processes and Strategies for the same queries using the same processes, it shows that the process contribution and the strategy priority have a significant impact on accurately determining the significance level of queries.

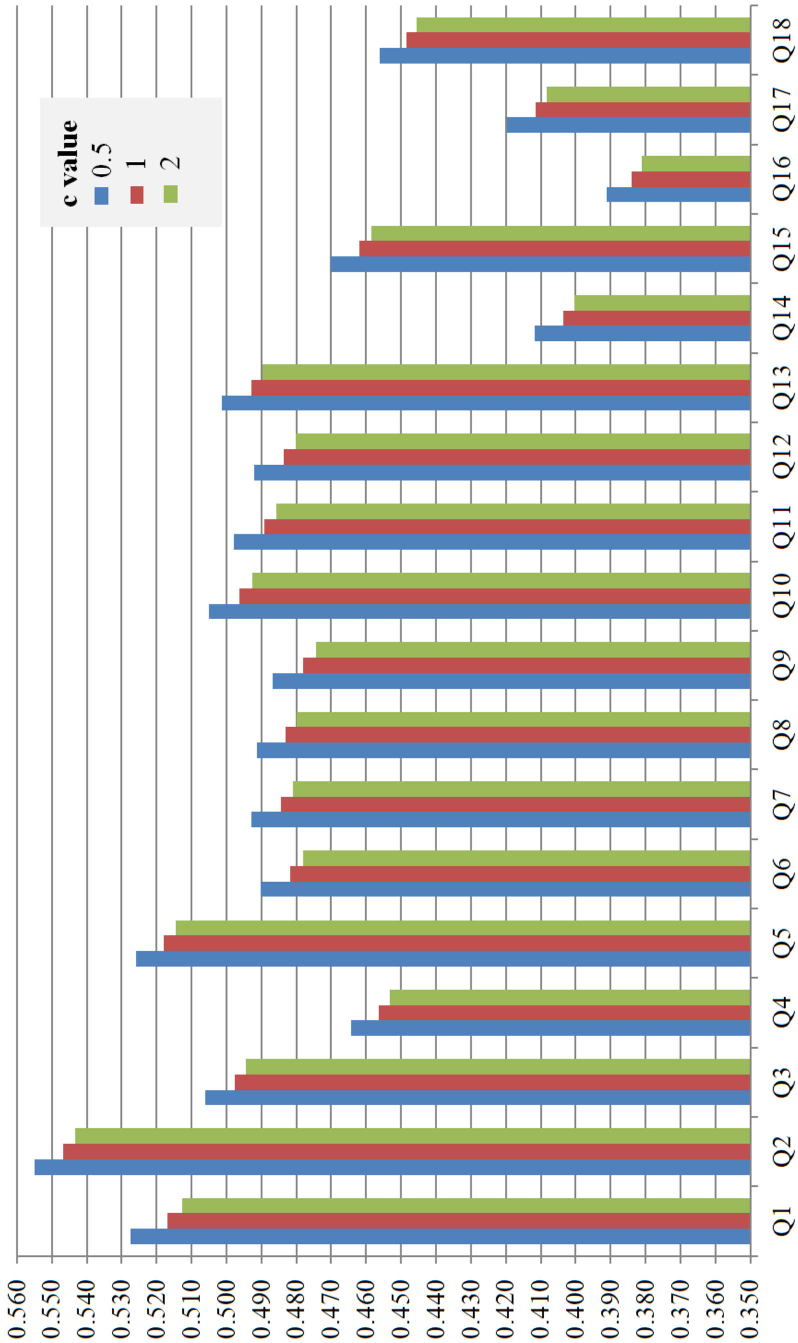

Figure 4.5: Significance levels of queries

In Significance Level of a Query Considering Business Processes, the significance level of a query was calculated based on the contribution weight of a process which was only defined by two values, i.e., 0.5 (non-core) and 1 (core). Whereas, the contribution weight of a process in Significance Level of a Query Considering Business Processes and Strategies is not narrowed within those two values, but it is a continuous value within [0,1]. This captures the process contribution more accurately than that of Significance Level of a Query Considering Business Processes. Furthermore, Significance Level of a Query Considering Business Processes and Strategies also exploits the contribution of a process to strategies and the strategy priority directly. Therefore, it is expected that the significance level of a query reflects the importance of a query for a business organisation more accurately.

4.3 Conclusions

In the approach 'Significance level of a query considering business processes', we have emphasized the importance of the significance level of a query to a business organisation for data collection and analysis in order to capture the deep insights of data. For the first time, we have proposed an approach to determine the significance level considering the processes of an enterprise system as a business case. The results have revealed that queries with higher semantic similarity values to the core business processes have higher significance levels. By mathematically quantifying the significance level of a query for a business organisation, an enterprise system can assess how important a query is, and it can then process the query using the appropriate scale of semantic information for both data collection and analysis. This will also improve the precision and computational complexity of information retrieval and data collection compared with processes where semantic information is not considered.

In this chapter, two approaches, namely 'Significance level of a query considering business processes' and 'Significance level of a query considering business processes and business strategies' have been introduced. The significance levels reflect not only the importance of a query to business processes but also emphasises the relevance between queries and business strategies. The results produced by both approaches based on a case study have demonstrated that, the more relevance a query has, the higher its significance level. Therefore, the enterprise information system in a business organisation can understand which query is more important to focus on and spend more time to extract deep insights into data. Therefore, this can improve the efficiency and effectiveness of an information system in exploiting big data for better decision-making and the adoption of new evaluations of current business strategies. Thus far, we have introduced all approaches required to determine the significance level of a big data query. The significance level of a number of complex queries has been evaluated based on a case study of a business organisation. However, the impact of the significance level of a query has not been assessed on the textual big data collection and analytics. The next chapter presents an approach to big data collection and analytics based on the significance level of a query and business contextual information.

Chapter 5

Semantic Information Scaling and Business Context for Big Data Analytics

As stated in Chapter 2, big data are generated from a variety of sources with different representation forms and formats. This raises a research question as to how "important data" on information relevant to a business context and semantically meaningful to a textual query can be captured and analysed more accurately to obtain deep and relevant business insights. As stated in Chapter 2, a number of big data analytic methods are available which consider contextual information such as the context of a query and its users, and the context of a query-driven recommendation system. However, these methods still have many challenges and none has considered the context of a business and the scaling of semantic information to a query in either data collection or analysis. To address this research gap, which was formulated as Objective 3 in Section 1.3 of Chapter 1, a big data analytical technique which embeds the information semantically meaningful to a textual query in terms of the significance level of a query and business contextual information into the bedrock of its data collection and analysis process is proposed in this chapter. The proposed model has been implemented under the framework of Hadoop considering a textual big data query relevant to a grocery shop as a case. The results exhibit that the approach substantially increases the amount of data collection and their deep insight with an increase of the significance level value and the extent of business contextual information.

5.1 An Overview of the Proposed Approach

Big data have a wide variety of types, including text, image, audio and video, and are continuously generated from diverse sources, such as sensor devices, networks, the web, social media, mobile applications and data storage. With the explosive increase of global data, many enterprises, not only large but also medium- and small-sized, have recently become aware of the importance of capturing key information from big data. This can be seen not only in the way they are investing their money in buying new technologies to extract precious information from big data but also in accelerating big data research and applications [167].

As the consideration of context and its relevance in big data analysis has consistently been the focus of the data science community, a number of existing data analytic methods consider contextual information including, but not limited to, information retrieval considering the context of a query and its users, query-driven context aware recommendation and so on, as mentioned in Chapter 2. However, these systems still have many challenges and to date none has considered the context of a business in either data collection or analysis. In contrast, it is clearly evident from the research in big data analytics, that if the business context and the semantic information on query keywords are embedded into both data collection and analysis, it is possible to capture more and deeper and more accurate insights into the data relevant to the business. This will help a business in the development of a feasible and achievable business strategy. To address this important research issue, for the first time, a big data analytic approach which considers the business context and a query's semantic words in both data collection and analysis is proposed in this chapter.

A fundamental concept of our proposed approach is presented in the system flowchart shown in Figure 5.1. As shown in this figure, after receiving a query, the approach processes the query based on five main processing phases as listed below:

(i) determine the significance level of the query, collect query semantic keywords and calculate the semantic values of these keywords;

(ii) extract semantic keywords from the list of keywords collected from phase (i) using the significance level of the query;

(iii) collect data based on the semantic keywords extracted from phase (ii) and the business context;

(iv) analyse data collected from phase (iii) using the keywords from phase (ii) and the business context from phase (iii);

(v) export the final expected outcome, i.e., business intelligence with the reports and visualisation of the results obtained from phase (iv).

In addition, the flowchart consists of three inputs:

(i) annotations of business processes and facts on business strategies for the first phase;

(ii) priorities of business strategies for the first phase;

(iii) business context for the third phase.

Figure 5.1 shows the main processing phases and their sequential order and inputs/outputs. To understand the approach more clearly, the next section describes the approach in detail.

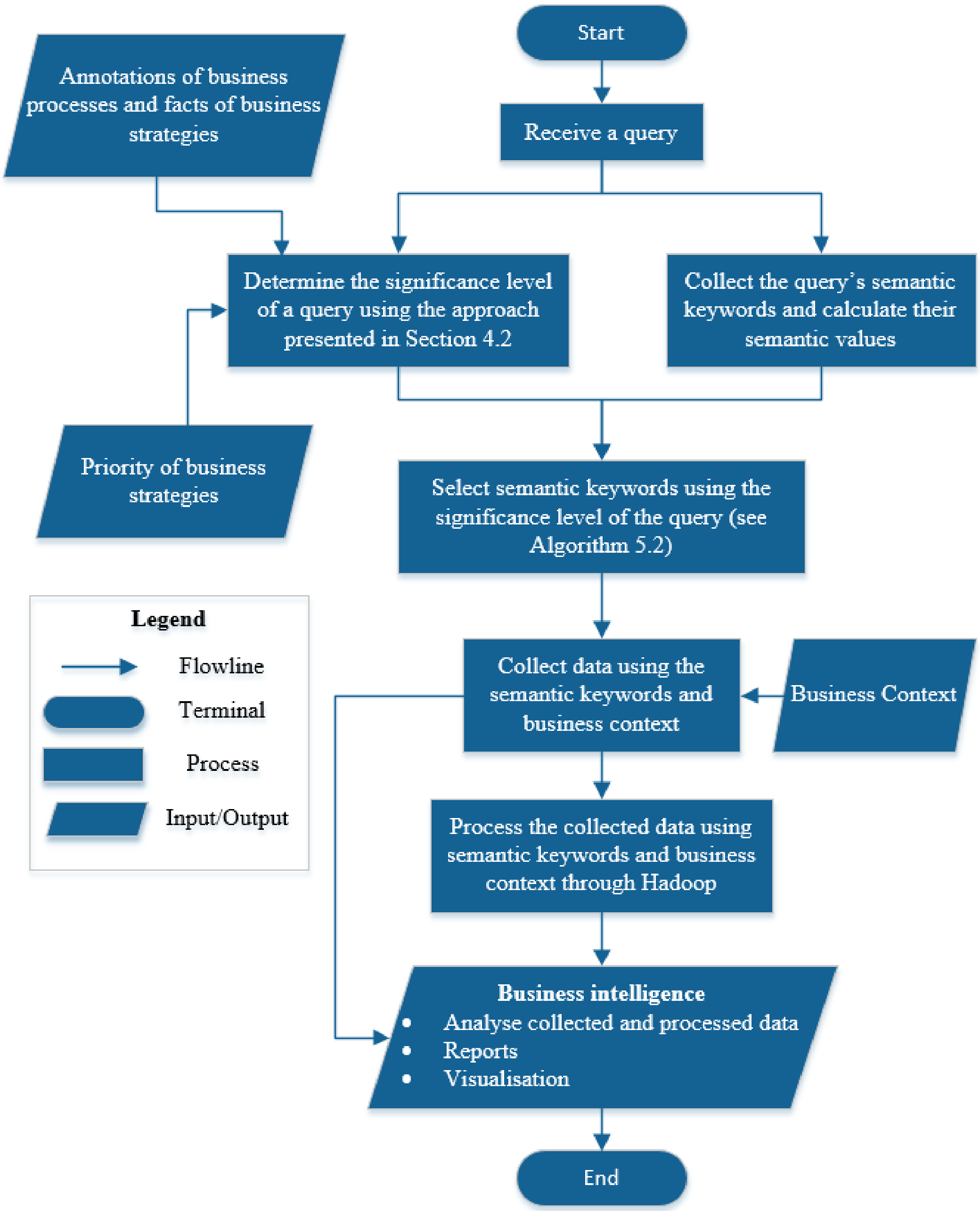

Figure 5.1: System flowchart of proposed approach

5.2 Description of Proposed Approach

The schematic diagram of the proposed approach is presented in Figure 5.2. In this figure, there are nine blocks and each block represents a process or an input or output. The first block (Block 5.2.1) is an input query, and the description of a textual query is detailed in the next section. The query is then analysed in the second block (Block 5.2.2) to collect its semantic keywords and calculate their semantic values. The next block (Block 5.2.3) is to determine the significance level of the input query using the results from Block 5.2.2. Based on the significance level, a list of selected semantic keywords is extracted, as shown in Block 5.2.4. These selected semantic keywords and the business context obtained from Block 5.2.7 are used in the next block (Block 5.2.5) to collect data from the data sources in Block 5.2.6. Block 5.2.8 highlights the process of analysing the data collected from Block 5.2.5 for the input query using the selected query semantic keywords received from Block 5.2.4 and the business context from Block 5.2.7 through Hadoop. Finally, Block 5.2.9 presents the results that are produced from Block 5.2.8, either in reports or visual form. Each block is explained in the following sections.

5.2.1 Type of queries used in proposed approach

The term “query” itself has a broad meaning. It could be a request for data collection or simply an inquiry for specific information. For example, users can put their queries into the database system to ask for information on loyal customers, or a manager can create a query to request the quarterly revenue, or a query from a customer asking for a refund.

In the present, we focus on the query for data collection and analysis. Although a data query is normally related to structured data such as queries in

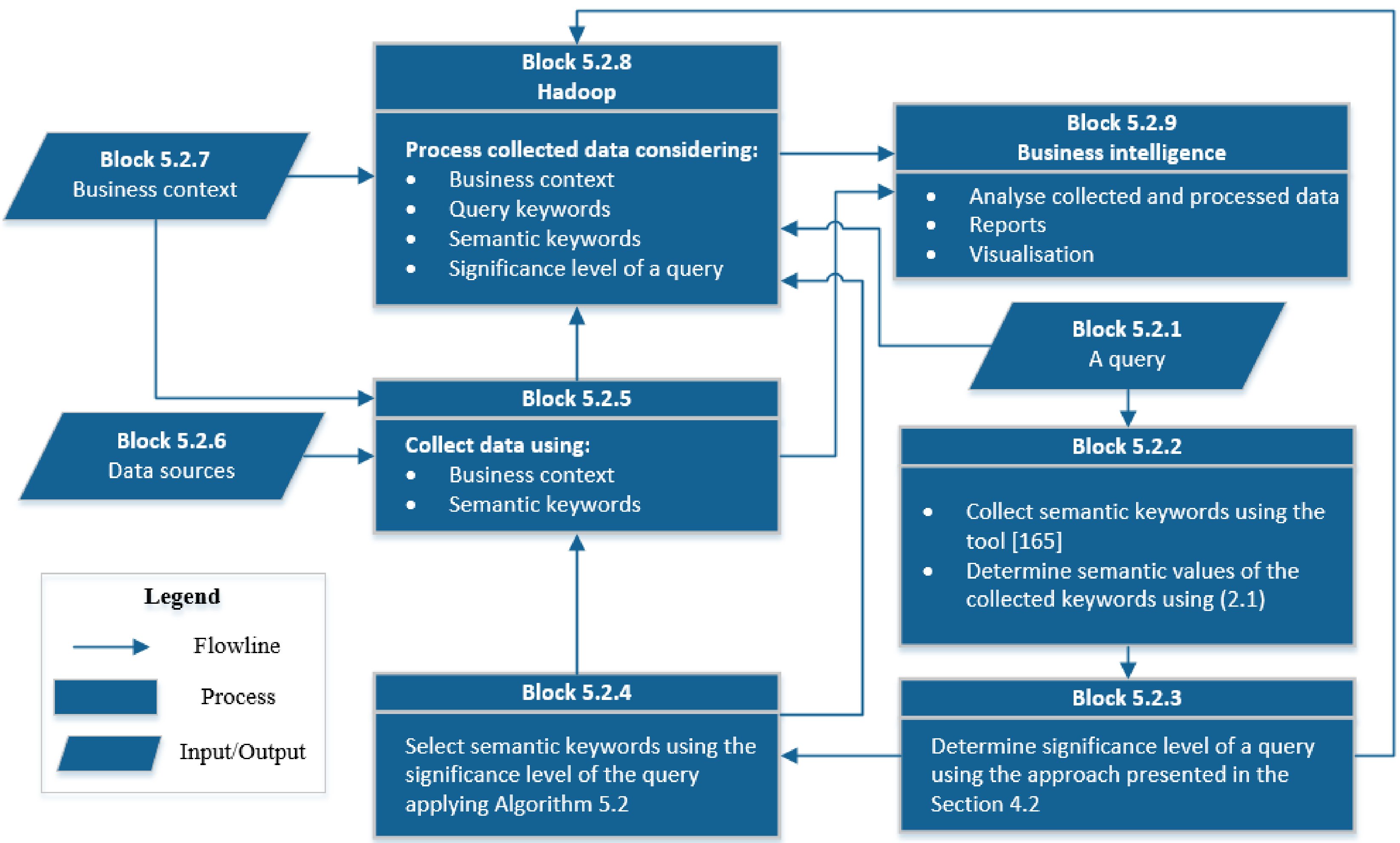

Figure 5.2: Schematic diagram of our proposed big data analytic approach considering business context and a query's semantic keywords

Structured Query Language (SQL), the concept of a query language for unstructured data has been released since the 1990s [168]. However, the study in [168] only considered how to deal with structured data that lack information in several of their entities by adopting the labelled tree structure in their data system.

As data query processing is still limited and only effective for structured data, here we present a new perspective on the analysis of queries for unstructured data. Therefore, our aim in this research is to consider only queries for unstructured data, specifically for textual data.

As discussed in previous chapters, many different textual queries are currently in use by business organisations. A list of sample queries that can be used in our proposed approach is shown in Table 5.2.1. However, the business scenario considered in this chapter is still a grocery store as considered in previous chapters and the focus is on the queries for the area: (i) fruit and (ii) vegetables. Therefore, the query keywords for the queries in this area are "fruit" and "vegetable".

5.2.2 Query semantic keywords and their semantic values

The semantic words of the query keywords, are found by using the tool Semantic Link [165]. This tool returns a set of semantic words that most frequently occurs with the given keyword using English Wikipedia. For instance, for the keyword "fruit", the tool returns "vegetable", "apricot", "desserts" and "diets". If more than one query keywords is found in a query, the lists of semantic words from these query keywords are then combined in one list.

Table 5.1: Semantic words and their semantic values in case study of "fruit" and "vegetable"

No	Semantic word	Semantic value	No	Semantic word	Semantic value
1	fruit	1	51	oranges	0.55
2	vegetable	1	52	raisins	0.55
3	acorns	0.82	53	soybeans	0.55
4	asparagus	0.82	54	sugarcane	0.55
5	berries	0.82	55	watermelon	0.55
6	cucumber	0.82	56	cheeses	0.54
7	herbs	0.82	57	meat	0.54
8	mushrooms	0.82	58	pasta	0.54
9	olives	0.82	59	seafood	0.54
10	onions	0.82	60	crops	0.48
11	seeds	0.82	61	beef	0.45
12	spinach	0.82	62	breads	0.45
13	almonds	0.67	63	cassava	0.45
14	apples	0.67	64	cereals	0.45
15	apricot	0.67	65	chilli	0.45
16	banana	0.67	66	coconuts	0.45
17	beans	0.67	67	mutton	0.45
18	broccoli	0.67	68	noodles	0.45
19	cabbage	0.67	69	pork	0.45
20	carrot	0.67	70	sausages	0.45
21	cherries	0.67	71	ghee	0.37
22	citrus	0.67	72	maize	0.37
23	figs	0.67	73	poultry	0.37
24	flowers	0.67	74	sepals	0.37

Continued on next page

Table 5.1 – continued from previous page

No	Semantic word	Semantic value	No	Semantic word	Semantic value
25	grains	0.67	75	wheat	0.37
26	grapes	0.67	76	basket	0.3
27	lettuce	0.67	77	carrion	0.3
28	mango	0.67	78	foodstuffs	0.3
29	melon	0.67	79	garlic	0.3
30	nuts	0.67	80	sorghum	0.3
31	ovary	0.67	81	diets	0.2
32	parsley	0.67	82	dishes	0.2
33	peaches	0.67	83	farmlands	0.2
34	pear	0.67	84	juices	0.2
35	Peas	0.67	85	jute	0.2
36	peppers	0.67	86	soup	0.2
37	pineapple	0.67	87	vitamins	0.2
38	pomegranate	0.67	88	gravy	0.16
39	potato	0.67	89	salad	0.16
40	pulses	0.67	90	stew	0.16
41	savory	0.67	91	tofu	0.16
42	sprouts	0.67	92	broth	0.14
43	stamens	0.67	93	greenhouses	0.14
44	strawberries	0.67	94	spices	0.14
45	tomato	0.67	95	sauce	0.11
46	buds	0.55	96	cacao	0.09
47	grapefruit	0.55	97	pickles	0.09
48	inflorescences	0.55	98	mayonnaise	0.07
49	lemons	0.55	99	dispersal	0.06

Continued on next page

Table 5.1 – continued from previous page

No	Semantic word	Semantic value	No	Semantic word	Semantic value
50	limes	0.55	100	cuisines	0.05

The linguistic semantic aspect is considered in our proposed approach, depending on the semantic similarity between the semantic words and the query keywords that can be derived using a lexical database. The semantic similarities between the semantic words and the query keywords are calculated using 2.1 in Chapter 2. The semantic values of the semantic words "fruit" and "vegetable" are listed in Table 5.1. In this table, because "fruit" and "vegetable" are also query keywords, their semantic values are 1.

As a semantic word may be associated with many query keywords with different semantic values, the final semantic value of this semantic word is the maximum value of their semantic values. For example, the semantic value between "herb" and "fruit" is 0.17, whereas the semantic value between "herb" and "vegetable" is 0.82. The final semantic value between "herb" and the query keywords is max(0.17, 0.82) where the max function returns the maximum value of the given arguments.

Several semantic words, such as "seafood" and "meat" seem to be not relevant to the query keywords, but their semantic values are high (0.54 each), whereas, the semantic word "salads" that has close semantic meaning to "vegetable" and "fruit" has the lower semantic value of 0.16. This is because the distance between "vegetable" and "seafood" is 3, while the distance between "vegetable" and "salads" is 9 in the lexical language system, as shown in Figure 5.3.

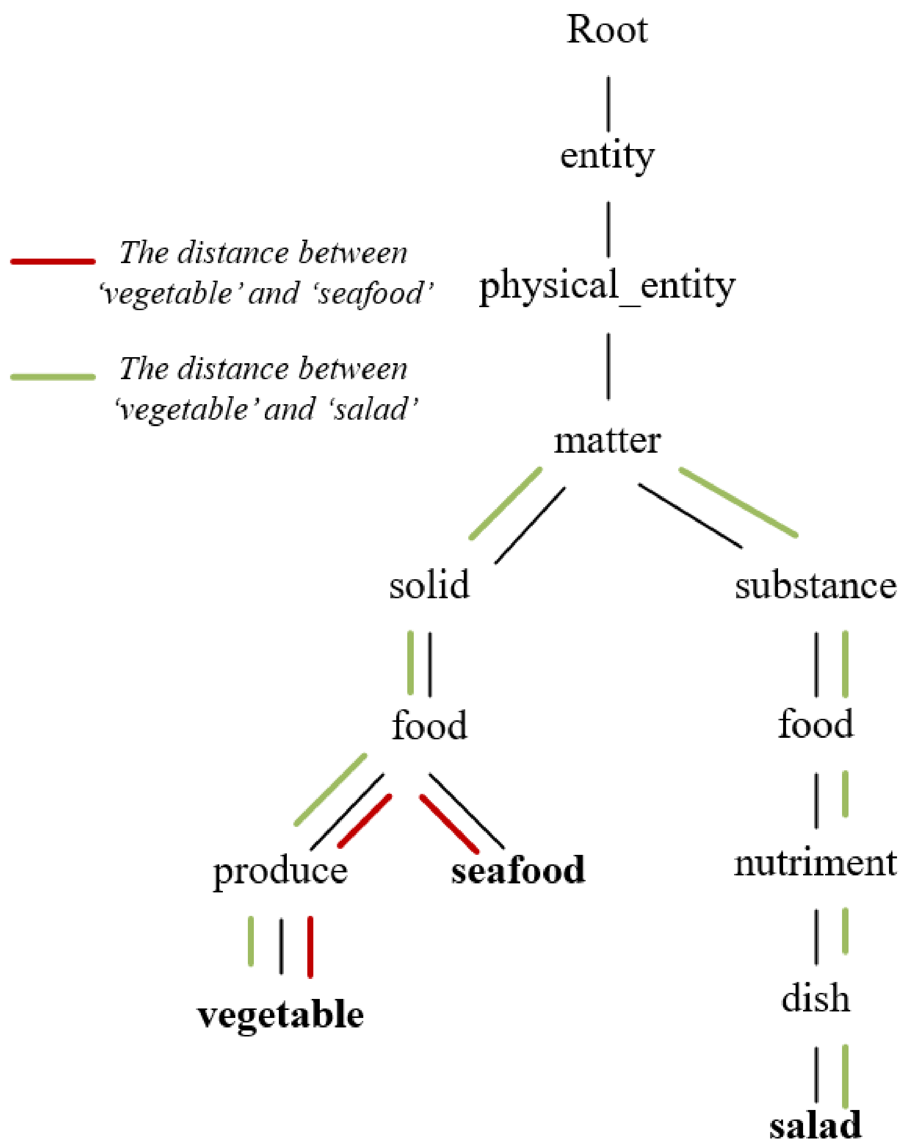

Figure 5.3: Distance between "vegetable" and "seafood" compared with distance between "vegetable" and "salad" in lexical language system

Furthermore, in the lexical language system, the words in the upper layers are more general and hence less detailed than the words in the lower layers. Therefore, the semantic similarities between the words in upper layers are also less than those of the word in lower layers [159]. Consequently, as the hierarchy distance from the root to the subsumer of the pair - "vegetable" and "seafood" is 6 and thus longer than the hierarchy distance of the subsumer of the pair - "vegetable" and "salad" (4), "seafood" has a higher semantic similarity than "salad". This indicates that the semantic interpretation using the lexical hierarchy may not always conform to human semantic perceptions.

5.2.3 Significance level of query

The significance level of a query is calculated as explained in Section 4.2 of Chapter 4 . In Chapter 4, two approaches were introduced. We selected the second approach namely 'Determine Significance Level of Query Considering BP and BS' to determine the significance level of the query in order to reflect business strategies and business processes more accurately, as explained in this chapter.

5.2.4 Select query semantic keywords

The method of selection of query semantic keywords is detailed in Algorithm 5.2. In this algorithm, all the semantic words are listed in descending order of their values (Step 1 of Algorithm 5.2). The input significance levels are used to select the semantic keywords from their position on the list after the query keywords. Based on the significance level of the query, a list of selected keywords and the query keywords are extracted in Step 2 of the algorithm.

For example, the context of a grocery shop is selected and the query is "How were the customers' experiences with fruit and vegetables purchased from the shop?". Note that for the above query, the query keywords are "fruit" and "vegetables". The semantic words of these query keywords and their semantic values are automatically determined as mentioned above in Table 5.1. 100 words have semantic relations to the query keywords (fruit, vegetables). These keywords are sorted in descending order by their semantic values (Step 1 of Algorithm 5.2) while keeping the query keywords at the top of the list.

Next, to apply the value of the significance level using Step 2 of the Algorithm 5.2, the approach selects the required keywords. The approach selects a different number of keywords from the top of the list for the different values of the significance level. For example, if the significance level equals 1.0, all 100 keywords are selected, while for the significance level 0.4, the first $\lceil 0.4 \times 98 \rceil + 2) = 42$ keywords in the list are selected, where $\lceil \rceil$ is a ceiling function.

Algorithm 5.2 Select semantic words

Input: The significance level (s) of the query and a list of semantic words including query keywords and their semantic values.
Ouput: A list of selected words

1. Sort list of all words in descending order of their semantic values but keep the query keywords at the top of the list.
2. Select the words from the list.

```
IF S = 1 THEN
    Select all the words in the list.
ELSE IF S = 0 THEN
    Select only the query keywords.
ELSE
    Select the query keywords and ⌈ S × (size of the list - n) ⌉ of the
    words from the list after the position of the query keywords where
    n represents the number of query keywords.
ENDIF
```

5.2.5 Data collection using selected query semantic keywords and business context

Here, business context is represented by the keywords that represent the various activities and constituent parts of a business. This is explained later in Section 5.2.7. To integrate the business context in a query in this study, each query possesses a significance level which indicates its importance for a business. The significance level determines the number of semantic keywords for a query keyword(s) that need to be considered in both data collection and analysis. To represent the importance of a keyword for a business context, each keyword has a semantic value, which is determined considering both the business and linguistic semantic aspects.

As mentioned in Chapter 4, the significance level of a query can be determined depending on how relevant or important it is to the business. Chapter 4 also showed which criteria a significance level of a query can be affected and how to calculate them. The significance level of a query is in the range of [0,1], and the more important queries have higher significance levels. The significance level of the query helps to scale data collection and analytics in terms of obtaining more relevant data to discover important clues.

5.2.6 Data sources

The data sources used in this research were public domains such as productreview.com.au and au.trustpilot.com, as shown in Figures 5.4 and 5.5 for customer reviews of thousands of brands (e.g., Woolworths, Aldi, Coles, Dyson, Miele). We selected customer reviews of a supermarket to analyse the customer feedback in this research. Since the data collected from these sources are text reviews from customers, these data are in unstructured format.

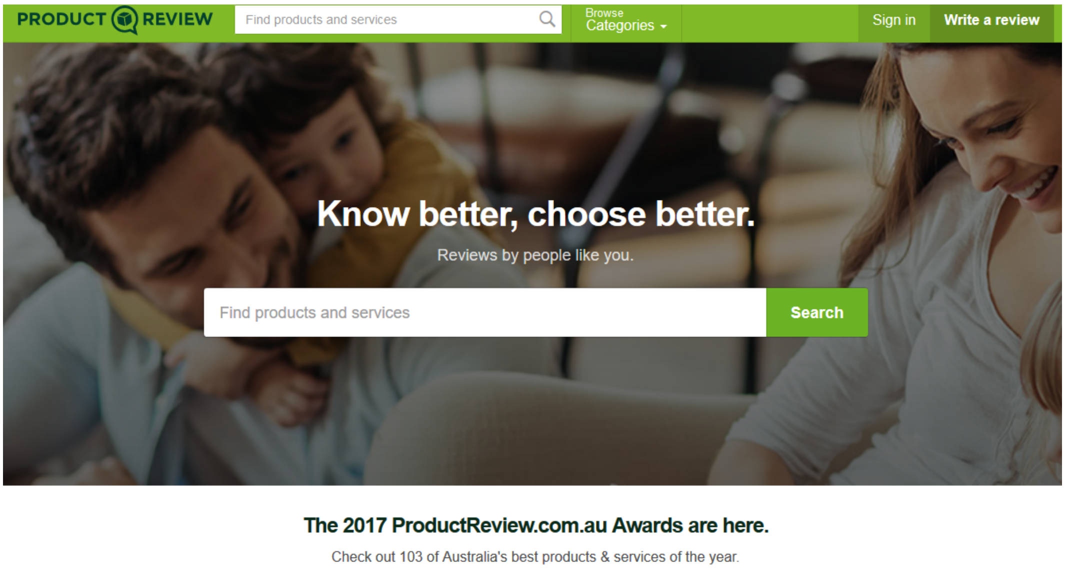

Figure 5.4: Website of Product Review (productreview.com.au)

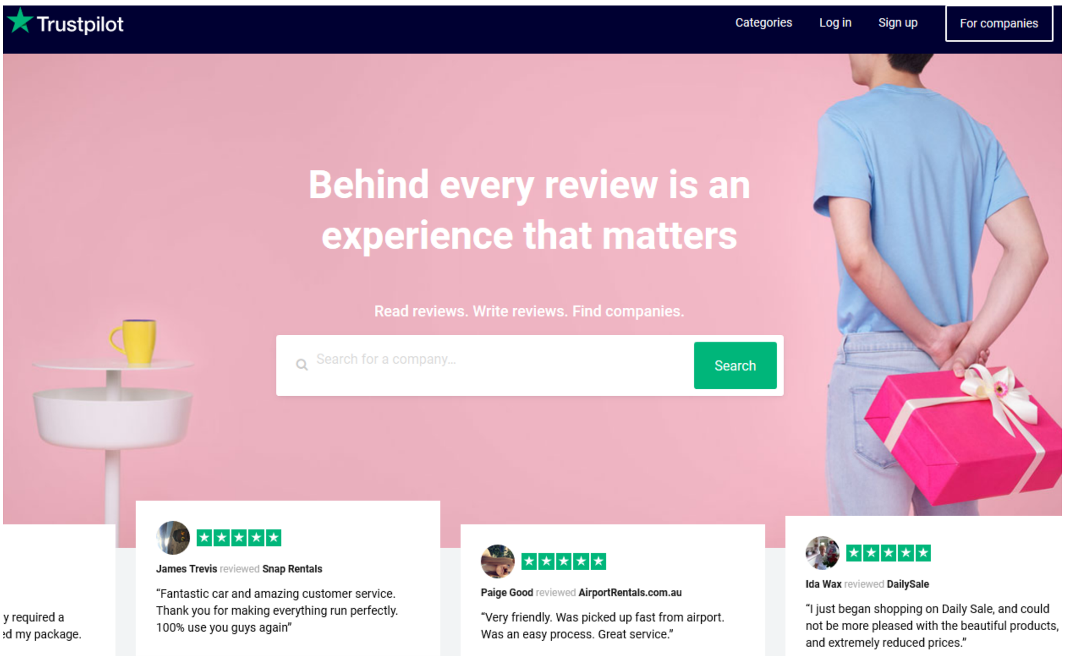

Figure 5.5: Website of Trustpilot (au.trustpilot.com)

5.2.7 Business context representation

While the linguistic semantic aspect represents the semantic interpretation based on the linguistic meaning such as synonyms and hypernyms, business organisations mainly focus on the quality of their products, services and facilities. This provides rich experience and satisfaction with their customers and puts them on better strategic position compared with other similar competitive business organisations; consequently, this helps a business organisation build solid finance foundation, and sustainable and successful business plan.

The business organisations use many ways to convey their messages related to their business strategy directly to their customers including slogans, advertisements online, in newspapers, on radio and TV, and distributing catalogues and leaflets. These types of messages to their customers are useful in representing the business context associated with their business strategies.

Since there is no definition of a context, anybody can use any useful and relevant information to represent a business context. In the present study, we use message information related to the slogans of a business to represent its context and the keywords of the slogans to represent the business context. This is because slogans convey the most important messages to customers.

Table 5.2: Companies and their slogans

Company	**Slogan**
Hungry Jacks	The Burgers Are Better At Hungry Jacks
Mc Donald's	I'm lovin' it
KFC	Finger Lickin' Good Nobody does chicken like KFC

Continued on next page

Table 5.2 – continued from previous page

Company	Slogan
	So good
Coles	Down down
Aldi	Good different
IGA	It pays to shop independent
Qantas Airways	The spirit of Australia
Eva Air	Sharing the world, flying together
Qatar Airways	Going places together
Apple	Think different
Samsung	Imagine Do what you can't
Dell	Yours is here Purely you

Table 5.3: Slogans for "fruit" and "vegetables" and their messages

Slogan	Message send to customers	Keywords for business context
Australia's fresh food people	Product quality	Australia, fresh, food, people
Famous dishes start with fresh ingredients	Product quality	famous, dishes, start, fresh, ingredients
If it's not fresh it's free	Customer guarantee. (If customers are not happy, they can get their money)	not fresh, free
Fresh or free	Customer guarantee.	fresh, free, guarantee

Continued on next page

Table 5.3 – continued from previous page

Slogan	Message send to customers	Keywords for business context
guarantee	(If customers are not happy, they can get their money)	
Fresh cut, ready to cook	Product quality and convenience	fresh cut, ready, cook
Aussie grown	Australian grown	Aussie, grown
The future of fresh	Product quality	future, fresh
Get into mango season with 100% Aussie mangoes	Australian grown	mango, season, 100%, Aussie
Fresh in for the weekend	Product quality	fresh, weekend

Here, the words used for business slogans are considered the aspects which most highlight the business. This is because a slogan refers to a promise that a company makes to its customers [169]. It is also a short sentence that makes it easy for customers to remember the benefits when they use a product. Furthermore, a slogan is very close to the goals and strategies a company need to reach.

Selected slogans are listed in Table 5.2 to show the messages that the companies which use these slogans want to send to their customers. As Hungry Jacks, McDonald's and KFC are fast food companies, they focus on the quality and taste of the food that offer to their customers. Each keyword in their slogans is totally focused on promoting their product, such as "burger", "better", "love", "finger", "chicken" and "good".

Other companies relevant to the retail market in Australia, such as Coles, Aldi and IGA concentrate on both their products and prices. The keyword of Coles' slogan is "down". This makes their customers associate with the lower price of products that Coles can offer to them. In contrast, Aldi reminds their customers that Aldi has "the difference" compared to the other retail providers that customers can always find products with the lowest prices in their stores. Similarly, IGA also wants to attract their customers' attention by highlighting the price via keywords in their slogan (e.g., pay, shop, independent).

In the field of aviation, while Qantas Airways emphasises the nationalism of their services by using keywords such as "spirit" and "Australia", other airlines such as Eva Air focus on the comfortable and friendly environment that they provide to their passengers by using keywords such as "sharing", "world", "flying" and "together". Qatar Airways defines their brand as one of the best and biggest airlines in the world by the keyword "places".

Technology companies all focus on the exclusive features of their products. For example, Apple emphasises the creativity of their product innovation by using the keywords "think" and "different". In contrast, Samsung concentrates on the exclusive features that their products can bring to their customers by using the keywords "do" and "can't", and Dell allows their customers to customize a desktop or laptop based on the customer's needs by using the keywords "purely" and "you".

A company may have more than one slogan. For example, KFC has three slogans, and Samsung and Dell have two slogans, as shown in Table 5.2. This is because slogans are developed depending on the features of products, the strategies of companies for their products and their marketing campaigns. For a query about "fruit" and "vegetables", a list of slogans, their messages sent to customers and possible keywords representing a business context relevant to fruit and vegetables for a retail company are shown in Table 5.3. For example, if the slogan is "Australia's fresh food people", the message sent to customers is about product quality (the quality of food) and the possible keywords that represent the business context are "Australia", "fresh", "food" and "people".

Table 5.4: Contextual keywords and their values

Contextual keyword	Semantic value in business context
fresh	1
Aussie	1
free	0.8
grown	0.8
guarantee	0.7
ingredient	0.6
cook	0.4
cut	0.2
weekend	0.2

Several contextual keywords selected from Table 5.3, with their contextual values for a business context assigned intuitively are given in Table 5.4. These words are then added to the list of query semantic words in Table 5.1 in the descending order of semantic value.

5.2.8 Big data processing with Hadoop

The keywords of a query are used to process the information directly relevant to the query. In addition, to collect deep insight about business processes relevant to the query, we use semantic information that is represented by the query's semantic keywords selected by the approach described in Section 5.2.4.

To capture the information directly associated with a business organisation (the slogans of a business organisation are used in this study), we also use business contextual keywords. For this study, the main function of Hadoop in data processing is to:

- select the query's semantic keywords using the significance level calculated following the approach described in Chapter 4.
- filter the data file of customer reviewers using all the keywords (query keywords, query's semantic keywords, business context keywords), as described above.

The implementation of the above mentioned functionalities in the Hadoop framework is detailed as follows:

In the Hadoop environment, a program is executed based on the concept of the Map-Reduce paradigm. In the Map-Reduce paradigm, key-value pairs are the heart of the process of Hadoop. The map phase performs the filtering and sorting of data based on the keys of the key-value pairs, while the reduce phase performs a summary operation based on the values of the key-value pairs [170]. Although the Map-Reduce programming paradigm is effective for huge amounts of data processing, to make the data analysis suit what users

Algorithm 5.3 Algorithm for Hadoop's Mapper Filter Function (MFF)

Input:

- The data file of customer reviews
- The list of semantic keywords
- The list including query keywords and business context keywords.
- The value of significance level

Output: List of reviews that contain at least one semantic keyword or query keyword or business context keywords.

1. Select semantic keywords and query keywords from the list of semantic keywords
using the given significance level value and Algorithm 5.2.
2. Generate a pattern using the selected keywords (semantic keywords and query keywords)
and the business context keywords.
3. Return the customer reviews from the data file that match the pattern.

need, user defined functions are required to provide the personalised functions for both the map and reduce phases.

For this study, our program mainly required Map phase, because the Reduce phase is mainly for the purpose of grouping data. Therefore, since we needed to filter customer reviews with query and business contextual keywords and the semantic words of the query, the Reduce phase was not necessary in this case.

To implement the required mapping function for this research project, we developed a Java class file called MFF (Mapper Filter Function) in Hadoop's development environment. The implementation of this research project was completed by following three main steps, as shown in Figure 5.6: (i) producing the MFF file in Java using Algorithm 5.3 , (ii) uploading the MFF file to the Pig script in the Hadoop environment, (iii) the results (customer reviews) selected and then returned by Hadoop were analysed based on user's needs using a sentiment analysis technique (http://semantria.com/), and represented by the report and visualisation parts of Business Intelligence. The source code of the Pig script required for Step ii is described below and shown in Figure 5.7.

In the above source code, first, the jar file is registered. Next, the significance levels (SL) contained in the file named ”significance_levels” are loaded in the variable “A”. All the comments are then loaded in the variable “B” from a data file on customer reviews called “comments” using TextLoader. Each comment is considered as a paragraph and defined as chararray. Fourthly, for each paragraph loaded, if it contains any word belonging to the pattern generated by the function, namely the “key” defined in MFF file, that paragraph is returned by the function and stored in the variable “C”. Similarly, all the paragraphs that follow the pattern match are stored in C. Finally, C is saved in a file called “result.csv”. The file “results.csv” is then analysed, as shown in Step 3 in Figure 5.6.

5.2.9 Business intelligence (BI)

For the purposes of the present study, Business Intelligence (BI) comprises three main blocks, as shown in Figure 5.8. Block 5.8.1 is the input data which are returned by Hadoop as explained in Section 5.2.5. After the data are transferred to the BI tool, they are analysed (see Block 5.8.2 of Figure 5.8) using a semantic text analysis tool produced by Semantria (https://semantria.com/) to determine the customers experience (e.g., negative, neutral, positive). Finally, the results produced in Block 5.8.2 are presented as reports, figures and charts on dashboards, and scorecards and reported using presentation and visualisation tools. As our focus in this research was not on BI, we introduce this section to show the effect of significance level and business context on the final outcome of BI process.

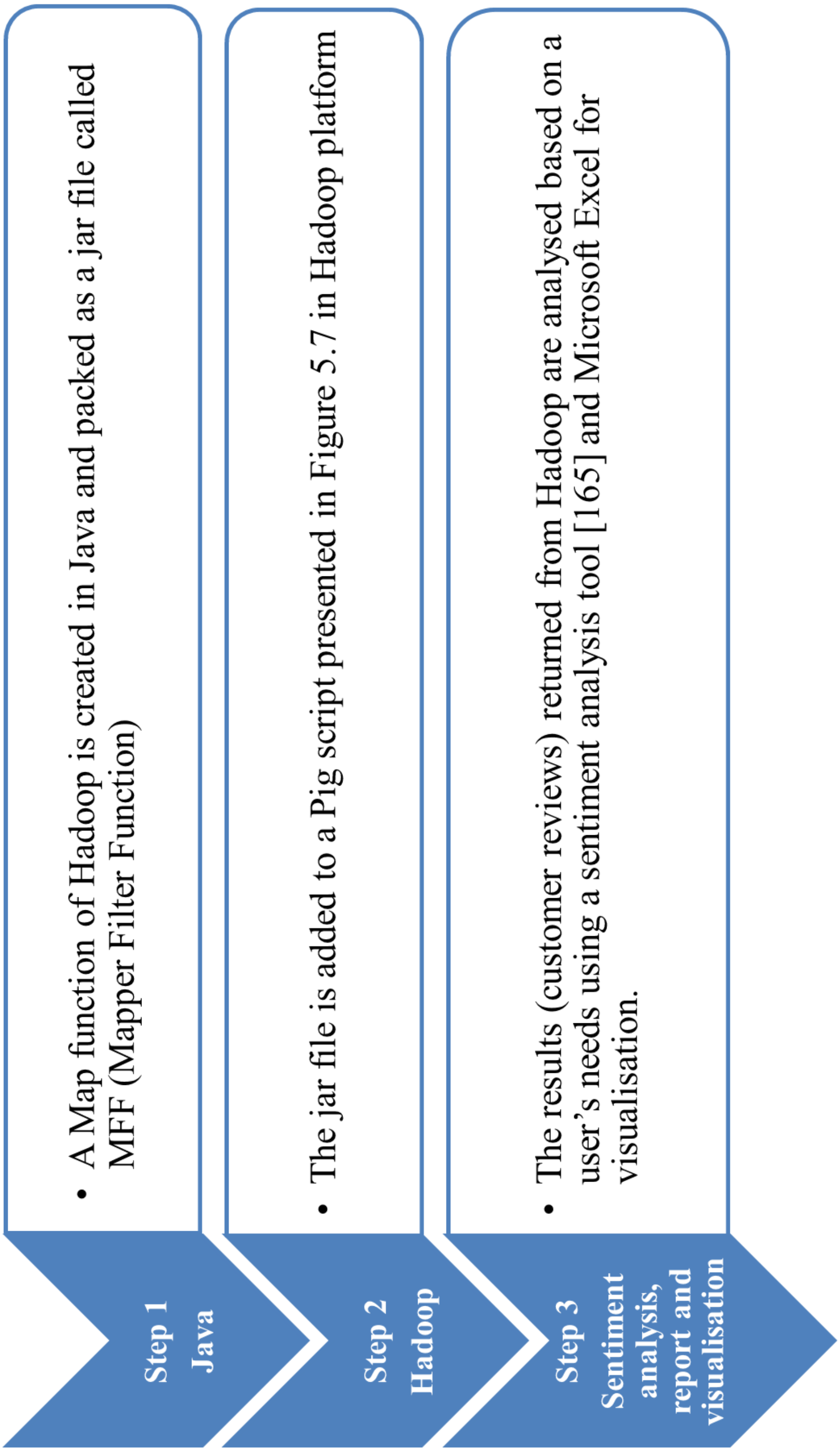

Figure 5.6: Brief description of implementation in Hadoop's development environment

```
REGISTER MFF.jar
-- The main function of  MFF.jar is key(input text, semantic keyword list, business context keywords,
significance level, 1)
A=LOAD 'significance_levels' As (SL: float);
-- 'significance_levels' is the file that contains the value of significance levels and already uploaded in
HBase
B= LOAD 'comments' USING TextLoader As (comment:  chararray);
-- 'comments' is the file that contains all customer reviews and already uploaded in HBase
C= FOREACH A GENERATE {
        FOREACH B GENERATE MFF.key(comment,'semantic_words.csv','business_words.csv',SL,1)
};
STORE C INTO 'results';
-- The results produced in this program are stored in the file 'results' which is located in HBase
```

Figure 5.7: Source code of Pig script developed for our approach

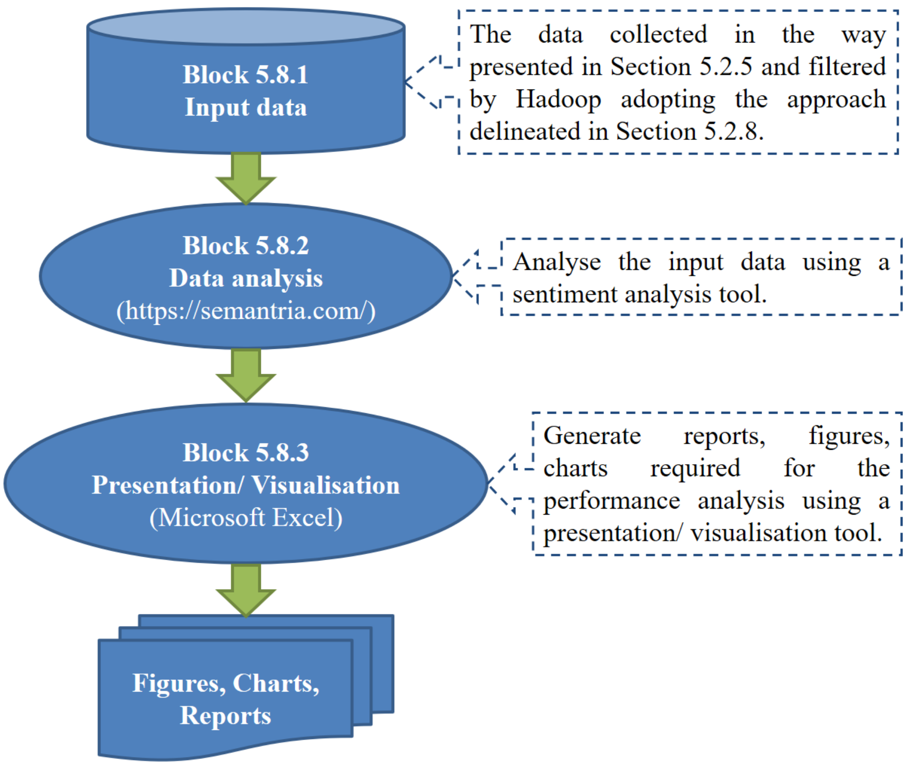

Figure 5.8: Main processes of business intelligence used in this study

5.2.9.1 Analysis of collected data

Of the contextual and semantic keywords, most customers feedback was collected for the “fresh” and “apple” keywords, i.e., 320 and 56 for “fresh” and “apple”, respectively. Since a large amount of customers’ feedback was acquired using the query semantic keywords and business contextual keywords, respectively, this indicates the benefits of using these keywords in terms of capturing more relevant information (here customers’ feedback).

The distribution of the collected data on customer experience for significance level = 0.2 using the approach articulated in Section 5.2.5 is shown in Table 5.5. In this table, the first two words “fruit” and “vegetable” in the query semantic keywords section are query keywords and the other words in this section are semantic keywords. The table shows the results for a low significance value (0.2) and two contextual keywords “fresh” and “aussie”, show a large amount of customer experience (116, 104, 120 feedback for positive, neutral and negative, respectively) which is 28.6% of the total 1188 of feedback items. This vindicates the rationale of using business context to capture more relevant data during data collection. Although a majority of customers feedback experience (614 out of 1188) was captured using the query keywords “fruit” and “vegetable”, still 234 i.e., 19.70% more customers experience was captured using the semantic keywords “fruit” and “vegetable”.

Table 5.5: Distribution of customer experience across selected keywords for significance level = 0.2

Type of keywords	Keyword	#Positive	#Neutral	#Negative	Total
Contextual	fresh	110	98	112	320

Continued on next page

Table 5.5 – continued from previous page

Type of keywords	Keyword	#Positive	#Neutral	#Negative	Total
keywords	aussie	6	6	8	20
Query semantic keywords	fruit	154	121	177	452
	vegetable	48	50	64	162
	acorns	1	0	0	1
	asparagus	0	2	0	2
	berries	0	6	14	20
	cucumber	0	1	6	7
	herbs	0	4	4	8
	mushrooms	0	2	6	8
	olives	6	6	4	16
	onions	2	8	10	20
	seeds	0	0	4	4
	spinach	2	4	8	14
	almonds	4	0	0	4
	apples	14	16	26	56
	apricot	0	0	2	2
	banana	10	10	12	32
	beans	2	2	10	14
	broccoli	2	4	8	14
	cabbage	2	0	1	3
	carrot	1	0	8	9

The data collected using the different values of significance level are presented in Figure 5.9. This figure shows that with an increase of significance level, the numbers of customer feedback relevant to the query and the set of semantic and contextual keywords in all types of experience increase. However, the rate of increase for neutral feedback is much less compared with the growth rates of positive and negative feedback. This is because customers normally tend to give their feedback when they are either extremely happy with the products and services or disappointed with the services or the products that they have experienced. In addition, according to the data collected, customers are not happy with fruits and vegetables from the selected retailer. Consequently, the number of negative experiences is the highest proportion in total. Note that the contextual keywords shown in Table 5.4 were intuitively selected in proportion to the value of the significance level.

For example, for a query without considering the significance level, i.e. significance level=0, the proposed approach captured 220 customer experience (negative=100, neutral=70 and positive=50), while it captured 560 (negative=270, neutral=120 and positive=170) for the significance level of 1.0. This indicates the capturing of more and deeper insights into customer experience by our proposed approach.

5.2.9.2 Analysis of data processed by Hadoop

Thus far we have investigated the influence of significance level and business context on data collection. Now, we present the impact of significance level and business context on data filtered by Hadoop with the patterns produced using the three types (query, semantic and context) of keywords. To show the impact, we calculated the relevance of data filtered by different values of significance levels and different extents of business context. Three different sets of business context were used, as shown in Figure 5.10.

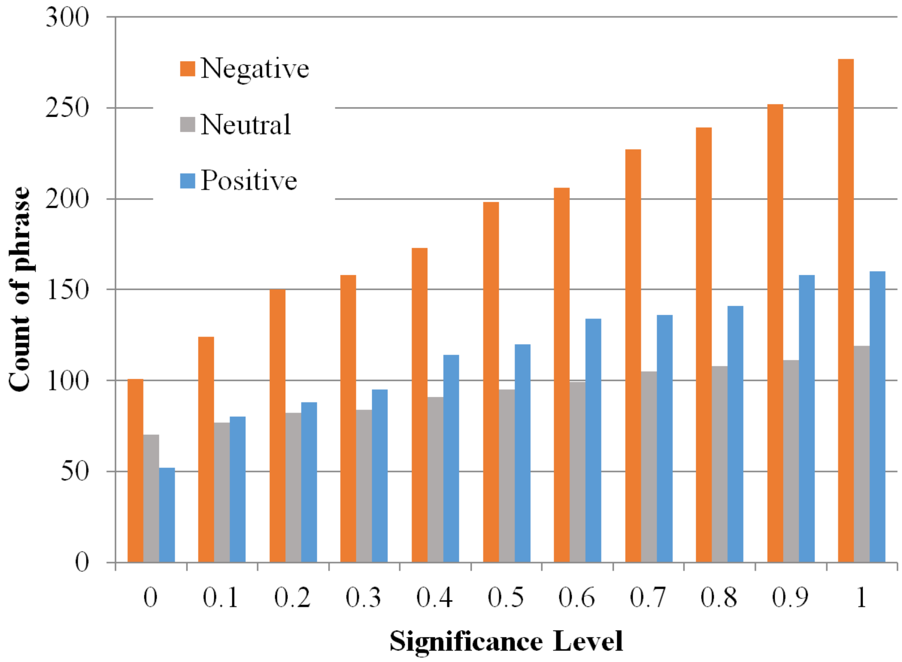

Figure 5.9: Number of negative, neutral, positive customer feedbacks relevant to "fruit" and "vegetables" for each significance level

Figure 5.10 shows the impact between the query significance levels and business contexts following data filtering with Hadoop. It can be seen that the more context words and the higher the value of significance levels considered during data filtering, the more relevant data are captured. This reveals that the data selected without contextual keywords are the lowest for all values of significance level. Moreover, by adding nine contextual keywords during data processing, the amount of data captured increases remarkably for all significance levels. With a significance level of 0.1, the portion of relevant data collected without context is less than 0.4, whereas it is almost 0.6 by adding nine context keywords. Similarly, with a significance level of 1, the relevance ratio without using context is slightly higher than 0.6 while it is more than 0.9 for nine contextual keywords. In addition, the relevance scores when the significance level is 1 for the other two cases, i.e., with three and six context keywords, are 0.83 and 0.88, respectively. These scores are also considerably higher than those for the case without using context.

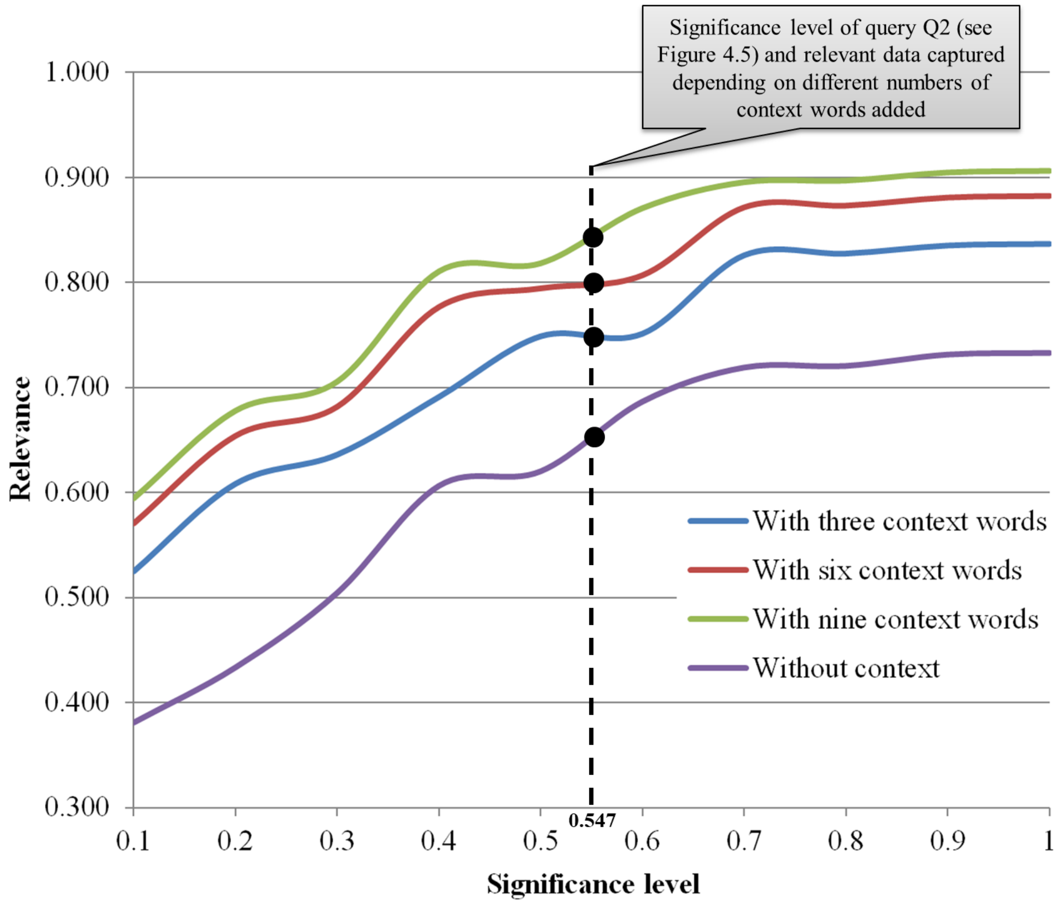

Figure 5.10: Impact of query significance level and business context on relevance of data collected by proposed approach

Furthermore, according to the approach described in Section 4.2 in Chapter 4 the value of the significance level of query Q2 is 0.547. For this value, in Figure 5.10, we also observe considerable improvement in the relevance data captured by Hadoop for the consideration of business contextual information.

5.3 Conclusions

In this chapter, we propose a new concept in big data analytics. By considering the business context during big data processing, not only deep insights of big data are captured but also more relevant and important information from these unstructured big data is collected. Unlike traditional big data analytic methods that focus on the keywords of a query, our method also considers the business context, depending on the significance level of a query, as well as the semantic keywords relevant to the query.

The business context presented in this chapter reflects the business strategic plan from product, service and brand aspects. The semantic keywords relevant to a query are generated dynamically considering the keywords of the query. We also scale the data collection and data processing techniques based on the semantic values of query keywords and business context keywords. We implement our method in Hadoop platform using the context of a retail shop using different significance values, and contextual keywords. The results exhibit an increased amount of data captured and deep insights in terms of customers' experience with an increase of significance level and business contextual information.

Although we present a complete textual big data analytic technique by exploiting business context and semantic information associated with a text-based query, there is some scope to further improve the study presented in this thesis. Concluding remarks on the achievements and findings of this study and possible future extensions are presented in the next chapter.

Chapter 6

Conclusions and Future Research

6.1 Conclusions

The significance of big data for a business for informed decision-making is increasing day by day. Since big data is massive and can come from any source in any structure, its complex nature has led to the goal of transforming raw data into meaningful information becoming a huge research challenge. Numerous studies have proposed a number of techniques to improve the effectiveness of big data analytics. However, because these techniques do not exploit business contextual information and the dynamic scaling of semantic information, their effectiveness is very limited.

To address this research issue and achieve the research objectives outlined in Chapter 1, we propose a technique for scaling semantic manipulation dynamically and integrating business contextual information in big data analytics. The key achievements of this thesis are summarised below:

- The rule-based inference model proposed in Chapter 3 describes how to establish the relationship between business processes and strategies via annotations and facts, respectively. A business strategy requires a number of business processes to be involved and a business process belongs to a number of business strategies. Each rule in the rule-based inference model contains facts which belong to business strategies and annotations belonging to business processes. Based on these rules, the links between business processes and strategies are automatically established. This rule-based inference model achieves the purpose of establishing a strong connection between business processes and their relevant strategies which will help enterprises to assess and hence reach their goals. This is also an integrated part of the approaches for estimating the significance level of a query, which are presented next.

- Two innovative approaches are proposed in Chapter 4 to determine the significance levels of queries. The first approach calculates the significance level of queries based on the intuitively selected weight of business processes, whereas, the second approach determines the significance level of queries, depending on the contributions of business processes and the priorities of business strategies. Because of the consideration of process contribution and business strategy priority, the second approach appears to be a better way to calculate the significance level of queries reflecting the business perspectives. This is because the latter approach dynamically considers the business function and strategic direction directly. The purpose of calculating the significance level of queries is to apply it in dynamically scaling the use of semantic interpretation in both data collection and analytics based on their importance in fulfilling business strategies.

- Finally, a technique is proposed to embed business context and dynamic scaling of semantically meaningful information in big data analytics considering the significance level of a textual big data query presented previously (see Chapter 6). Based on the significance level of queries, the use of semantic interpretation in both data collection and processing is dynamically scaled. The results produced using this technique show that for queries with higher significance levels, more relevant data and deeper insights are captured. The same thing was observed for the consideration of the business context, i.e., the higher the exposure of business contextual information, the more relevant the data collected and the deeper the insights obtained in data analytics.

6.2 Future Research

As with other research projects, the approaches proposed in this thesis can be improved in many different ways. Some are described as follows:

1. Once a business process is contextualised, it can be used in data mining for targeted collection of data and then processed more accurately by applying semantic interpretation. This can be used in collecting more data but with increased relevancy through semantic manipulation, thereby enabling the capture of deep insights. In this thesis, we have introduced a method that addresses this issue for text type data only, but it is not suitable for other data types (e.g., video, audio, image). Hence, it remains a challenge to model the business process to enable big data to handle diverse data types concurrently. Capturing semantic information and business context from these other data types suitable to a query requires advanced content-based analysis.

2. In addition to considering slogans as business contextual information, many other types of information are useful for developing a business contextual information model. For example, information on competitors with similar businesses, information on the economic and political context that may affect the revenue of a business, such as economic crisis, unemployment rate and elections. The competitors information can be captured by analysing their

advertisements, new product campaigns, promotions and their customers reviews.

3. As B2B collaborations are included in BPM, exchanging information/data between business organizations and their partners such as suppliers and retailers requires stringent privacy and a lightweight and adaptive security policy. Rapid technological development is gradually making communication and computing devices very smart and tiny. As all of these devices are connected through the pervasive computing environment, they are always connected and business services are available on any device at any time from anywhere across the world. This demands a security framework with pervasive security mechanisms (e.g., intrusion detection, security audit trails) for BPM/WFM, the enforcement of strict privacy among businesses and partners and a scalable and adaptive security policy.

Appendix

Publications based PhD Research

.1 Articles that have been published

Peer reviewed International Conference Papers:

1. Loan Thi Ngoc Dinh, Gour Karmakar, Joarder Kamruzzaman, and Andrew Stranieri. 'Business context in big data analytics,' in *10th International Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1-5. IEEE, 2015.(CORE ranking: B)

2. Loan Thi Ngoc Dinh, Gour Karmakar, Joarder Kamruzzaman, and Andrew Stranieri. 'A Rule-based Inference Model to Establish Strategy-Process Relationship,' in *30th International Business Information Management Association (IBIMA)*, 2017.(CORE ranking: B)

3. Loan Thi Ngoc Dinh, Gour Karmakar, Joarder Kamruzzaman, Andrew Stranieri, and Rajkumar Das. 'Significance Level of a Query for Enterprise Data,' in *30th International Business Information Management Association (IBIMA)*, 2017.(CORE ranking: B)

4. Loan Thi Ngoc Dinh, Gour Karmakar, Joarder Kamruzzaman, and Andrew Stranieri. 'Significance Level of a Big Data Query by Exploiting Business Processes and Strategies,' in *13th International Baltic Conference on Databases and Information Systems (BalticDBIS2018)*, 2018.(CORE ranking: B)

.2 Articles awaiting a decision

Journal Paper:

1. Loan Thi Ngoc Dinh, Gour Karmakar and Joarder Kamruzzaman, 'A Survey of Context Awareness in Big Data Analytics for Business Applications,' *Springer Journal of Knowledge and Information Systems*, 2018. submitted on 04 Jan 2018. (Impact factors: 2.247, CORE ranking: B)

Bibliography

[1] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *International conference on the practical applications of knowledge discovery and data mining*, 2000.

[2] E. Monk and B. Wagner, *Concepts in enterprise resource planning*. Cengage Learning, 2012.

[3] K. Li, H. Jiang, L. T. Yang, and A. Cuzzocrea, *Big data: algorithms, analytics, and applications*. CRC Press, 2015.

[4] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[5] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big data computing and clouds: trends and future directions," *Journal of Parallel and Distributed Computing*, vol. 79, pp. 3–15, 2015.

[6] M. F. Uddin and N. Gupta, "Seven v's of big data understanding big data to extract value," in *Zone1 Conference of the American Society for Engineering Education (ASEE Zone 1)*, 2014.

[7] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[8] J. Bai, J. Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[9] S. Smanchat, S. Ling, and M. Indrawan, "A survey on context-aware workflow adaptations," *Advances in Mobile Computing and Multimedia (MoMM)*, 2008.

[10] J. L. D. L. Vara, R. Ali, F. Dalpiaz, J. Sanchez, and P. Giorgini, "Business processes contextualization via context analysis," *Conceptual Modeling - ER*, vol. 6412, pp. 471–476, 2010.

[11] S. Boutanmina and R. Maamri, *A survey on context-aware workflow systems*. Security and Advanced Communication in Intelligent Information Processing, 2015.

[12] R. Aknouche, O. Asfari, F. Bentayeb, and O. Boussaid, *Integrating query context and user context in an information retrieval model based on expanded language modeling*. Berlin: Multidisciplinary Research and Practice for Information Systems, Springer, 2012.

[13] N. M. Adams, "Perspectives on data mining," *International Journal of Market Research*, vol. 52, no. 1, pp. 11–19, 2010.

[14] Z. Song and A. Kusiak, "Optimising product configurations with a data-mining approach," *International Journal of Production Research*, vol. 47, no. 7, pp. 1733–1751, 2009.

[15] X. Han, L. Tian, M. Yoon, and M. Lee, "A big data model supporting information recommendation in social networks," in *Cloud and Green Computing (CGC), 2012 Second International Conference on*. IEEE, 2012, pp. 810–813.

[16] R. A. Sinoara, J. Antunes, and S. O. Rezende, "Text mining and semantics: a systematic mapping study," *Journal of the Brazilian Computer Society*, vol. 23, no. 1, p. 9, 2017.

[17] D. Sánchez, M. J. Martín-Bautista, I. Blanco, and C. J. de la Torre, "Text knowledge mining: an alternative to text data mining," in *Data Mining*

Workshops, 2008. ICDMW'08. IEEE International Conference on. IEEE, 2008, pp. 664–672.

[18] L. H. Medida and K. Ramani, "Survey on semantic indexing of high dimensional data with deep learning techniques," *i-Manager's Journal on Software Engineering*, vol. 11, no. 2, p. 31, 2016.

[19] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in *Information communication and embedded systems (ICICES), 2013 international conference on*. IEEE, 2013, pp. 271–276.

[20] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, no. 2, pp. 293–314, 2014.

[21] R. Aknouche, O. Asfari, F. Bentayeb, and O. Boussaid, "Integrating query context and user context in an information retrieval model based on expanded language modeling," in *International Conference on Availability, Reliability, and Security*. Springer, 2012, pp. 244–258.

[22] A. Ghoting, P. Kambadur, E. Pednault, and R. Kannan, "Nimble: a toolkit for the implementation of parallel data mining and machine learning algorithms on mapreduce," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 334–342.

[23] L. Yu, J. Zheng, W. C. Shen, B. Wu, B. Wang, L. Qian, and B. R. Zhang, "Bc-pdm: data mining, social network analysis and text mining system based on cloud computing," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1496–1499.

[24] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen *et al.*, "Mllib: Machine learning in

apache spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[25] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed graphlab: a framework for machine learning and data mining in the cloud," *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[26] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[27] J. Akaichi, "Social networks' facebook'statutes updates mining for sentiment classification," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 886–891.

[28] R. Antai, "Sentiment classification using summaries: A comparative investigation of lexical and statistical approaches," in *Computer Science and Electronic Engineering Conference (CEEC), 2014 6th*. IEEE, 2014, pp. 154–159.

[29] F. Colace, M. De Santo, and L. Greco, "A probabilistic approach to tweets' sentiment classification," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 37–42.

[30] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, vol. 2. IEEE, 2013, pp. 91–96.

[31] L. Zhang and B. Liu, "Aspect and entity extraction for opinion mining," in *Data mining and knowledge discovery for big data*. Springer, 2014, pp. 1–40.

[32] N. Hariri, M. Bamshad, and B. Robin, "Query-driven context aware recommendation," in *ACM conference on Recommender systems*, 2013.

[33] A. Abbas, L. Zhang, and S. U. Khan, "A survey on context-aware recommender systems based on computational intelligence techniques," *Computing*, vol. 97, no. 7, pp. 667–690, 2015.

[34] S. Lee, S. Park, and S. G. Lee, *A study on issues in context-aware systems based on a survey and service scenarios*. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2009.

[35] D. Ejigu, M. Scuturici, and L. Brunie, *An ontology-based approach to context modelling and reasoning in pervasive computing*. Pervasive Computing and Communications Workshops, 2007.

[36] J. Xie, B. P. Knijnenburg, and H. Jin, "Location sharing privacy preference: Analysis and personalized recommendation," in *International Conference on Intelligent User Interfaces*, 2014.

[37] L. Sokol and S. Chan, "Context-based analytics in a big data world: Better decisions," *An IBM Redbooks Point-of-View publication*, 2013.

[38] L. T. N. Dinh, G. Karmakar, J. Kamruzzaman, and A. Stranieri, "A rule based inference model to establish strategy-process relationship," in *30th International Conference of Business Information Management Association (IBIMA)*, 2017.

[39] ——, "Significance level of a query for enterprise data," in *30th International Conference of Business Information Management Association (IBIMA)*, 2017.

[40] L. T. N. Dinh, G. Karmakar, J. Kamruzzaman, A. Stranieri, and R. Das, "Significance level of a big data query by exploiting business processes and strategies," in *International Baltic Conference on Databases and Information Systems, 2018 13th International Conference on*. (BalticDBIS, 2018.

[41] L. T. N. Dinh, G. Karmakar, J. Kamruzzaman, and A. Stranieri, "Business context in big data analytics," in *Information, Communications and Signal*

Processing (ICICS), 2015 10th International Conference on. IEEE, 2015, pp. 1–5.

[42] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," *ACM sIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.

[43] G. D. Abowd, A. K. Dey, P. J. Brown, P. J. Davies, N. Smith, and P. Steggles, *Towards a Better Understanding of Context and Context-Awareness*, handheld and ubiquitous computing ed. Berlin: Springer, 1999.

[44] A. Lorentz, "With big data context is a big issue, april 2013. [online]," available: [Accessed 5 Feb 2018]. [Online]. Available: http://www.wired.com/insights/2013/04/with-big-data-context-is-a-big-issue/

[45] P. Kulkarni, S. Joshi, and M. S. Brown, "Chapter 4: Long live the king of big data: The context," *Big Data Analytics, Delhi, PHI Learning Private Limited*, pp. 50–66, 2016.

[46] G. Chen and D. Kotz, "A survey of context-aware mobile computing research," *Technical Report TR2000-381*, vol. 1, no. 2.1, p. 1, 2000.

[47] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 2, no. 4, pp. 263–277, 2007.

[48] T. Strang and C. Linnhoff-Popien, "A context modeling survey," *International Workshop on Advanced Context Modelling, Reasoning And Management*, 2004.

[49] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, "A survey of context modelling and reasoning techniques," *Pervasive and Mobile Computing*, vol. 6, no. 2, pp. 161–180, 2010.

[50] W. Liu, X. Li, and D. Huang, "A survey on context-awareness," *Computer Science and Service System (CSSS)*, 2011.

[51] P. Bellavista, A. Corradi, M. Fanelli, and L. Foschini, "A survey of context data distribution for mobile ubiquitous systems," *ACM Computing Surveys (CSUR)*, vol. 44, no. 4, pp. 1–45, 2012.

[52] G. George, M. R. Haas, and A. Pentland, "Big data and management," *Academy of Management Journal*, vol. 57, no. 2, pp. 321–326, 2014.

[53] T. Rout, M. R. Senapati, M. Garanayak, and S. K. Kamilla, "Big data and its applications: A review," in *International Conference on Electrical Electronics Signals, Communication and Optimization (EESCO)*, 2015.

[54] S. Mishra, V. Dhote, G. S. Prajapati, and J. P. Shukla, "Challenges in big data application: A review," *International Journal of Computer Applications*, vol. 121, no. 19, pp. 42–46, 2015.

[55] P. Russom, "Big data analytics," Tdwi Best Practices Report, Tech. Rep., 2011.

[56] A. Rajendra, "Big data computing," CRC Press, Tech. Rep., 2013.

[57] S. Sagiroglu and D. Sinanc, "Big data: a review," in *International Conference on Collaboration Technologies and Systems (CTS)*, 2013.

[58] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big data: the next frontier for innovation, competition and productivity*. Mckensey Global Institute, 2011.

[59] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts methods and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[60] B. R. Presentation, "The challenge of big data,," 2012, ventana Research, [Accessed 19 March 2018]. [Online]. Available: http://

www.ventanaresearch.com/uploadedFiles/Content/Landing_Pages/Ventana_Research_Big_Data_Benchmark_Research_Presentation.pdf

[61] A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H. a. Jacobsen, "Big bench: Towards an industry standard benchmark for big data analytics," in *The ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2013.

[62] N. Elgendy and A. Elragal, "Big data analytics: A literature review paper, advances in data mining," *Applications and Theoretical Aspects*, vol. 8557, pp. 214–227, 2014.

[63] J. Gantz and D. Reinsel, "The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, 2012.

[64] "Data equity: Unlocking the value of big data," Tech. Rep., 2012, [Accessed 18 Jan 2018]. [Online]. Available: http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf

[65] P. S. Tan, A. E. S. Goh, and S. S. G. Lee, *An ontology to support Context-Aware B2B services*. Services Computing, 2010.

[66] M. Leppanen, *A context-based enterprise ontology*. Business Information Systems, 2007.

[67] I. Kroschel, *On the notion of context for business process use*. ISSS/BPSC, 2010.

[68] P. J. Brown, J. D. Bovey, and X. Chen, "Context-aware applications: from the laboratory to the marketplace," *Personal Communications*, vol. 4, no. 5, pp. 58–64, 1997.

[69] K. Ploesser, M. Peleg, P. Soffer, M. Rosemann, and J. C. Recker, "Learning from context to improve business processes," *BPTrends*, vol. 6, no. 1, pp. 1–7, 2009.

[70] H. Cao, D. H. Hu, D. Shen, D. Jiang, J. T. Sun, E. Chen, and Q. Yang, "Context-aware query classification," in *International ACM SIGIR conference on Research and development in information retrieval*, 2009.

[71] A. K. Dey, "Understanding and using context," *Personal and ubiquitous computing*, vol. 5, no. 1, pp. 4–7, 2001.

[72] B. N. Schilit and M. M. Theimer, "Disseminating active map information to mobile hosts," *Network*, vol. 8, no. 5, pp. 22–32, 1994.

[73] P. J. Brown, "A framework for creating context-aware applications," *Electronic publishing*, vol. 9, no. 1, pp. 259–272, 1996.

[74] N. Ryan, J. Pascoe, and D. Morse, "Enhanced reality fieldwork: the context-aware archaeological assistant," *Bar International Series*, vol. 750, pp. 269–274, 1999.

[75] A. K. Dey, "Context-aware computing: The cyberdesk project," in *Spring Symposium on Intelligent Environments*, 1998.

[76] T. Winograd, "Architecture for context," *Human-Computer Interaction*, vol. 16, no. 2, pp. 401–419, 2001.

[77] J. Coutaz, J. L. Crowley, S. Dobson, and D. Garlan, "Context is key," *Communications of the ACM*, vol. 48, no. 3, pp. 49–53, 2005.

[78] "Big data - a new world of oppotunities," Available: [Accessed 15 April 2018], Tech. Rep., December 2012. [Online]. Available: http://www.nessi-europe.eu/Files/Private/NESSI_WhitePaper_BigData.pdf

[79] P. S. Tan, A. E. S. Goh, and S. S. G. Lee, "A context model to support b2b collaboration," *Enabling Context-Aware Web Services: Methods, Architectures, and Technologies, CRC Press*, pp. 243–271, 2010.

[80] P. S. Tan, S. S. G. Lee, A. E. S. Goh, and E. W. Lee, "Context-enabled b2b collaborations," in *International Conference on Services Computing (SCC)*, 2007.

[81] O. Saidani and S. Nurcan, "Towards context aware business process modeling," in *Workshop on Business Process Modeling Development and Support, CAiSE*, 2007.

[82] M. Rosemann, J. Recker, and C. Flender, "Contextualisation of business processes," *International Journal of Business Process Integration and Management*, vol. 3, no. 1, pp. 47–60, 2008.

[83] I. Ruthven, "Information retrieval in context," *Advanced topics in information retrieval*, vol. 33, pp. 187–207, 2011.

[84] P. B. G. K. Mostfaoui and A. Generic, "Framework for context-based distributed authorizations," in *International and Interdisciplinary Conference on Modeling and Using Context*. Berlin: Springer, 2003.

[85] R. Ali, F. Dalpiaz, and P. Giorgini, "A goal-based framework for contextual requirements modeling and analysis," *Requirements Engineering*, vol. 15, no. 4, pp. 439–458, 2010.

[86] M. Rosemann, "Proposals for future bpm reasearch directions," in *Asia Pacific Business Process Management*. Springer International Publishing, 2014, pp. 1–15.

[87] B. R. and N. V., "Context-aware business intelligence," *European Business Intelligence Summer School*, vol. 253, pp. 87–110, 2015.

[88] D. Hollingsworth and U. K. Hampshire, "Workflow management coalition documentation," in *Workflow Management Coalition*, 1993.

[89] J. Jung, "Mapping of business process models to workflow schemata: an example using memo-orgml and xpdl," 2004, university Koblenz-Landau.

[90] M. Wieland, P. Kaczmarczyk, and D. Nicklas, "Context integration for smart workflows," in *International Conference on Pervasive Computing and Communications (PerCom)*, 2008.

[91] F. Daneshgar, "Context-aware framework for erp," *in Encyclopedia of Information Science and Technology, IGI Global*, pp. 569–572, 2005.

[92] S. Meilin, Y. Guangxin, X. Yong, and W. Shangguang, "Workflow management systems: a survey," in *International Conference on Communication Technology Proceedings (ICCT)*, 1998.

[93] V. J. Brocke, S. Zelt, and T. Schmiedel, "On the role of context in business process management," *International Journal of Information Management*, 2015.

[94] L. Williams, "Testing overview and black-box testing techniques," Tech. Rep., 2006, [Accessed 31 April 2018]. [Online]. Available: agile.csc.ncsu.edu/SEMaterials/BlackBox.pdf

[95] Q. Chen and M. Hsu, "Inter-enterprise collaborative business process management," in *International Conference on Data Engineering*, 2001.

[96] P. Soffer and Y. Wand, "On the notion of soft-goals in business process modeling," *Business Process Management Journal*, vol. 11, no. 6, pp. 663–679, 2005.

[97] M. Mora, M. Gomez, J. Garrido, and F. C. Perez, "Engineering and management of it-based service systems," *Intelligent Systems Reference Library*, vol. 2014.

[98] D. M. Silva, R. M. Araujo, R. M. Santoro, and F. M. Pascual, "Defining context in a business process collaborative elicitation approach," in *International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2012.

[99] M. Rosemann and J. C. Recker, "Context-aware process design: Exploring the extrinsic drivers for process flexibility," in *International Conference on Advanced Information Systems Engineering*, 2006.

[100] R. K. Ko, S. S. Lee, and E. W. Lee, "Business process management (bpm) standards: a survey," *Business Process Management Journal*, vol. 15, no. 5, pp. 744–791, 2009.

[101] T. Roeser and E. M. Kern, "Surveys in business process management - a literature review," *Business Process Management Journal*, vol. 21, no. 3, pp. 692–718, 2015.

[102] X. Zhao and S. Mafuz, "Towards incorporating context awareness into business process management, world academy of science, engineering and technology," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, vol. 9, no. 12, pp. 3890–3897, 2015.

[103] T. Hori and K. Aizawa, "Capturing life-log and retrieval based on contexts," in *International Conference on Multimedia and Expo*, 2004.

[104] Z. H. Zhou, "Three perspectives on data mining," *Artificial Intelligence*, vol. 143, no. 1, pp. 139–146, 2003.

[105] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010.

[106] P. Kumar, S. Gopalan, and V. Sridhar, "Context enabled multi-cbr based recommendation engine for ecommerce," in *E-business Engineering*, 2005.

[107] N. Ratprasartporn, J. Po, A. Cakmak, S. Bani-Ahmah, and G. Ozsoyoglu, "Context-based literature digital collection search," *The International Journal on Very Large Data Bases*, vol. 18, no. 1, pp. 277–301, 2009.

[108] K. Ghag and K. Shah, "Comparative analysis of the techniques for sentiment analysis," in *International Conference on Advances in Technology and Engineering (ICATE)*, 2013.

[109] B. Dehning and V. J. Richardson, "Returns on investments in information technology: A research synthesis," *Journal of Information Systems*, vol. 16, no. 1, pp. 7–30, 2002.

[110] N. Melville, K. Kraemer, and V. Gurbaxani, "Information technology and organisational performance: An integrative model of it business value," *MIS Quarterly*, vol. 28, no. 2, pp. 283–322, 2004.

[111] R. O. V, *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. Wiley, 2009.

[112] M. Nycz and Z. Polkowski, "Business intelligence in a local government unit," in *Proceedings of Informing Science and IT Education Conference (InSITE)*, 2015.

[113] S. Chakravarthy, V. Krishnaprasad, V. Anwar, and S. K. Kim, "Composite events for active databases: sementics, contexts and detection," *VLDB*, vol. 94, pp. 606–617, 1994.

[114] H. Herbst, "Business rules in systems analysis: a meta-model and repository system," *Information Systems*, vol. 21, no. 2, pp. 147–166, 1996.

[115] L. T. Moss and S. Atre, *Business Intelligence Roadmap: The complete Project Life cycle for decision-support applications*. Addison-Wesley Professional, 2003.

[116] S. Rivest, Y. Bedard, M. j. Proulx, and M. Nadeau, "Solap technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 60, no. 1, pp. 17–33, 2005.

[117] T. M. Gronli, G. Ghinea, M. M. Younas, and J. Hansen, "Automatic configuration of mobile applications using context-aware cloud-based services," *Modeling and Processing for Next-Generation Big-Data Technologies*, vol. 4, pp. 367–383, 2015.

[118] B. A. A. Diallo, T. Badard, F. Hubert, and S. Daniel, "Context-based mobile geobi: enhancing business analysis with contextual metrics/statistics and context-based reasoning," *GeoInformatica*, vol. 18, no. 2, pp. 405–433, 2014.

[119] W. Hopken, M. Scheuringer, D. Linke, and M. Fuchs, "Context-based adaptation of ubiquitous web applications in tourism," in *Information and Communication Technologies in Tourism*, 2008.

[120] X. Wu, X. Xhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[121] R. K. Lomotey and R. Deters, "Towards knowledge discovery in big data," in *International Symposium on Service Oriented System Engineering (SOSE)*, 2014.

[122] J. Wust, C. Meyer, and H. Plattner, "Dac: Database application context analysis applied to enterprise applications," in *Australasian Computer Science Conference*, 2014.

[123] P. W. Liang and B. R. Dai, "Opinion mining on social media data," in *Mobile Data Management (MDM)*, 2013.

[124] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval, vol. 2, no. 1-2*, vol. 2, no. 1-2, pp. 1–135, 2008.

[125] C. S. Tucker and H. M. Kim, "Optimizing product configurations with a data mining approach," *Journal of Mechanical Design*, vol. 130, no. 4, p. 041103, 2008.

[126] D. Che, M. Safran, and Z. Peng, "From big data to big data mining: Challenges, issues and opportunities," *Database Systems for Advanced Applications*, 2013, springer Berlin Heidelberg.

[127] P. Lak, M. Sadat, C. J. Barrelet, M. Petitclerc, A. Miranskyy, C. Statchuk, and A. B. Bener, "Preliminary investigation on user interaction with ibm watson analytics," in *The 26th Annual International Conference on Computer Science and Software Engineering (CASCON)*, 2016.

[128] "Microsoft-developer network: Context awareness," Microsoft, Tech. Rep., [Accessed 15 Jan 2018]. [Online]. Available: https://msdn.microsoft.com/en-us/library/dn632192.aspx

[129] "Spark programming guide," [Accessed 20 Jun 2018], Tech. Rep. [Online]. Available: http://spark.apache.org/docs/2.1.1/programming-guide.html

[130] L. A. Steffenel, O. Flauzac, A. S. Charo, P. P. Barcelos, B. Stein, S. Nesmachnow, M. K. Pinheiro, and D. Diaz, "Per-mare: Adaptive deployment of mapreduce over pervasive grids," P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), Tech. Rep., 2013.

[131] G. W. Cassales, A. S. Charao, M. K. Pinheiro, C. Souveyet, and L. A. Steffenel, "Context-aware scheduling for apache hadoop over pervasive

environments," in *International Conference on Ambient Systems, Networks and Technologies (ANT)*, 2015.

[132] G. Cassales, A. Charao, M. Kirsch-Pinheiro, C. Souveyet, and L. Steffenel, "Bringing context to apache hadoop," in *Mobile Ubiquitous Computing*, 2014.

[133] K. Park, Y. Kim, and J. Chang, "Semantic reasoning with contextual ontologies on sensor cloud environment," *International Journal of Distributed Sensor Networks*, vol. 2014, 2014.

[134] Z. Khan, S. Kiani, and K. Soomro, "A framework for cloud-based context-aware information services for citizens in smart cities," *Journal of Cloud Computing*, vol. 3, no. 1, p. 1, 2014.

[135] C. Misale, M. Drocco, M. Aldinucci, and A. G. Tremblay, "Comparison of big data frameworks on a layered dataflow model," in *International Symposium on High-Level Parallel Programming and Applications (HLPP' 16)*, 2016.

[136] G. Hesse and M. Lorenz, "Conceptual survey on data stream processing systems," in *International Conference on In Parallel and Distributed Systems (ICPADS)*, 2015.

[137] ibm.com, Tech. Rep., 2014, [Accessed 20 May 2018]. [Online]. Available: https://www.ibm.com/developerworks/community/files/basic/anonymous/api/library/9c97dc1a-6f24-4217-b8ae-da9332ecc905/document/80bf2120-acd3-4439-b411-f4a81afb3e2c/media

[138] N. Taherimakhsousi and H. A. Moller, "Context-aware real-time video analytics," in *25th Annual International Conference on Computer Science and Software Engineering*, I. Corp, Ed., 2015.

[139] J. E. Gerow, J. B. Thatcher, and V. Grover, "Six types of it-business strategic alignment: An investigation of the constructs and their measurement," *European Journal of Information Systems*, vol. 24, no. 5, pp. 465–491, 2015.

[140] D. A. Almajali and Z. M. Dahalin, "Factors influencing it-business strategic alignment and sustainable competitive advantage: A structural equation modelling approach," *Communications of the IBIMA*, 2011.

[141] E. D. Morrison, A. K. Ghose, H. K. Dam, K. G. Hinge, and K. Hoesch-Klohe, "Strategic alignment of business processes," in *International Conference on Service-Oriented Computing*. Springer, 2011, pp. 9–21.

[142] B. C. Madu, "Vision: The relationship between a firm's strategy and business model," *Journal of behavioral studies in business*, vol. 6, p. 1, 2013.

[143] M. Rosemann and J. vom Brocke, "The six core elements of business process management," in *Handbook on business process management 1*. Springer, 2015, pp. 105–122.

[144] M. Grabowska, J. Krzywda, and D. Krzywda, "Relations between business model and business strategy," in *The 2015 WEI International Academic Conference Proceedings*, 2015.

[145] J. Freed, "And now for the pilotless passenger plane. the sydney morning herald. [online]," available: [Accessed 08 Jan. 2018]. [Online]. Available: http://www.smh.com.au/business/aviation/and-now-for-the-pilotless-passenger-plane-20150924-gjucza.html

[146] F. K. Chou, E. T. Wang, and F. Yang, "Realizing it strategic alignment and business performance: An integration of three perspectives." in *PACIS*, 2015, p. 179.

[147] M. Queiroz, "Mixed results in strategic it alignment research: a synthesis and empirical study," *European Journal of Information Systems*, vol. 26, no. 1, pp. 21–36, 2017.

[148] W. Wang, M. Indulska, and S. Sadiq, "Integrated modelling of business process models and business rules: a research agenda." ACIS, 2014.

[149] P. Delfmann, M. Steinhorst, H.-A. Dietrich, and J. Becker, "The generic model query language gmql–conceptual specification, implementation, and runtime evaluation," *Information Systems*, vol. 47, pp. 129–177, 2015.

[150] J. Becker, P. Delfmann, H.-A. Dietrich, M. Steinhorst, and M. Eggert, "Business process compliance checking–applying and evaluating a generic pattern matching approach for conceptual models in the financial sector," *Information Systems Frontiers*, vol. 18, no. 2, pp. 359–405, 2016.

[151] M. Fellmann, P. Delfmann, A. Koschmider, R. Laue, H. Leopold, and A. Schoknecht, "Semantic technology in business process modeling and analysis. part 1: Matching, modeling support, correctness and compliance." in *EMISA Forum*, vol. 35, no. 1, 2015, pp. 15–31.

[152] K. Hinge, A. Ghose, and G. Koliadis, "Process seer: A tool for semantic effect annotation of business process models," in *Enterprise Distributed Object Computing Conference, 2009. EDOC'09. IEEE International*. IEEE, 2009, pp. 54–63.

[153] G. Governatori, J. Hoffmann, S. Sadiq, and I. Weber, "Detecting regulatory compliance for business process models through semantic annotations," in *International Conference on Business Process Management*. Springer, 2008, pp. 5–17.

[154] M. Fellmann, O. Thomas, and B. Busch, "A query-driven approach for

checking the semantic correctness of ontology-based process representations," in *Business Information Systems*. Springer, 2011, pp. 62–73.

[155] G. Governatori and S. Shek, "Rule based business process compliance." in *RuleML (2)*, 2012.

[156] S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren, and D. Valerio, "A software reference architecture for semantic-aware big data systems," *Information and software technology*, vol. 90, pp. 75–92, 2017.

[157] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[158] G. O. Consortium, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[159] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138–1150, 2006.

[160] C. D. Francescomarino, M. Rospocher, C. Ghidini, and A. Valerio, "The role of semantic annotations in business process modelling," in *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, Sept 2014, pp. 181–189.

[161] D. M. Riehle, S. Jannaber, P. Delfmann, O. Thomas, and J. Becker, "Automatically annotating business process models with ontology concepts at design-time," in *International Conference on Conceptual Modeling*. Springer, 2017, pp. 177–186.

[162] R. Guizzardi and A. N. Reis, "A method to align goals and business processes," in *International Conference on Conceptual Modeling*. Springer, 2015, pp. 79–93.

[163] A. Nieto-Rodriguez. (2016) How to prioritize your companys projects. [Online]. Available: https://hbr.org/2016/12/how-to-prioritize-your-companys-projects

[164] D. Lidow. (2017) A better way to set strategic priorities. [Online]. Available: https://hbr.org/2017/02/a-better-way-to-set-strategic-priorities

[165] M. Pacula. Semantic-link. [Accessed 20 May 2018]. [Online]. Available: http://semantic-link.com/

[166] D. Niyato and E. Hossain, "A microeconomic model for hierarchical bandwidth sharing in dynamic spectrum access networks," *IEEE Transactions on Computers*, vol. 59, no. 7, pp. 865–877, 2010.

[167] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.

[168] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu, "A query language and optimization techniques for unstructured data," in *ACM SIGMOD Record*, vol. 25, no. 2. ACM, 1996, pp. 505–516.

[169] M. Dass, C. Kohli, P. Kumar, and S. Thomas, "A study of the antecedents of slogan liking," *Journal of Business Research*, vol. 67, no. 12, pp. 2504–2511, 2014.

[170] F. Gui, F. Zhang, Y. Ma, M. Liu, and W. Shen, "Social relation extraction of large-scale logistics network based on mapreduce," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2273–2277.